Discriminative Learning and Visual Interactive Behavior

Learning Techniques in Audio-Visual Information Processing (ICPR Tutorial)

Tony Jebara
Brian Clarkson
Sumit Basu
Alex Pentland

MIT Media Lab

Outline

-*Motivation*
  - *Learning Tasks and Paradigms*

-*Discriminative / Conditional Learning*
  - *Maximum Entropy Discrimination*

-*Latent Variables and Reversing Jensen's Inequality*
  - *CEM and Dual of EM*

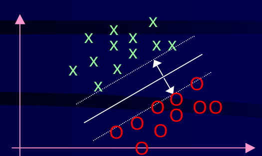
-*Action-Reaction Learning*
  - *Behavior Analysis / Synthesis via Time Series Prediction*

-*Wearable Platforms*
  - *Personal Enhanced Reality System*
  - *Wearable Interaction Learning*

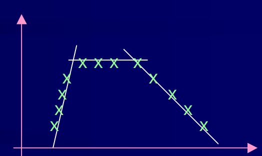
-*Conversational Context Learning*

Learning Applications for A & V

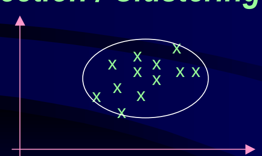
Classification:



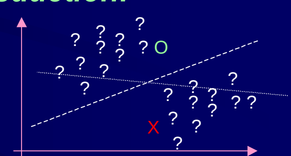
Regression / Prediction:



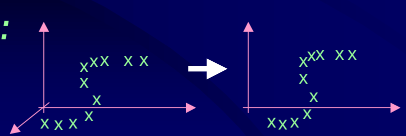
Detection / Clustering:



Transduction:



Feature Selection:



Learning Paradigms

New Competitors for Maximum Likelihood:
Discriminative Learning & SVMs *(NIPS, COLT, UAI, ICML)*

-*Time Series Prediction (Muller)*

-*Digit Recognition (Vapnik)*

-*Speech Recognition (Deng)*

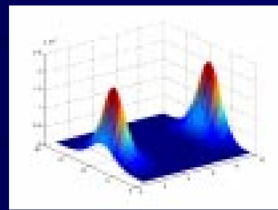
-*Face Gender Classification (Moghaddam)*

-*Gene-Sequence Classification (Jaakkola)*

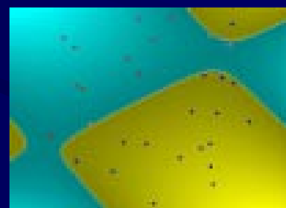
-*Text Classification (Joachims)*

Learning Paradigms

1) Generative Approach:
- Probabilistic Models
- Maximum Likelihood



2) Discriminative Approach:
- Support Vector Machines
- VC Dimension
- Maximum Margin
- Task is Explicit
- Discriminant Surface



Complementary Pros & Cons

1) ML & PDFs	2) Discrimination & SVMs
+Natural Models (HMMs)	+Model & Data Mismatch
+Priors	+Support Vectors
+Missing Data	+Good Generalization
+Flexible	-Linear Model
-Poor Performance	-Kernels
-Objective not Task Related	-No Priors
	-No Missing Data

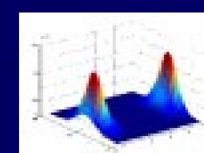
How to combine both? MED...

Maximum Entropy Discrimination

Tony Jebara
Tommi Jaakkola
Marina Meila

Overview & Motivation

Maximum Entropy Discrimination:
Combines probabilistic methods (and extensions) in discriminative framework
- -Add task-related Objective to PDFs.
- -Satisfy task constraints
- -Support Vectors -> Generalization
- -Convex, No Local Minima
  (vs. Min Class Error)



Feasible MED Extensions & Applications:

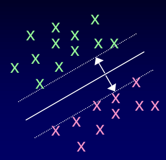
Latent variables, various priors, missing labels, structure estimation, anomaly detection.
Feature selection, regression, latent transformations, multi-class classification, exponential family.

Classification - Regularization Approach

Given: training examples: $\{X_1,\ldots,X_T\}$
binary (+/- 1) labels: $\{y_1,\ldots,y_T\}$
discriminant function: $L(X;\Theta)$
non-increasing margin loss: $l(.)$

Minimize: regularization penalty: $R(\Theta)+\sum_t l(\gamma_t)$
subject to classification constraints: $y_t L(X_t;\Theta)-\gamma_t \geq 0, \forall t$

Example: SVM



minimizes: $\frac{1}{2}\|\Theta\|^2+\sum_t f(\gamma_t)$
with discriminant: $L(X;\Theta)=\Theta^T X+b$
decision rule: $\hat{y}=sign(L(X;\Theta))$

Maximum Entropy Discrimination Approach

Many solutions may be valid.
Use coarser description of sol'n instead of a single optimum.
Solve for distribution P(Θ) over all good Θ (instead of Θ *).
Find $P(\Theta,\gamma)$ that mins $KL(P\|P_0)$ subject to constraints:

$$\int P(\Theta,\gamma)\left[y_t L(X_t;\Theta)-\gamma\right]d\Theta\, d\gamma \geq 0,\ \forall t$$

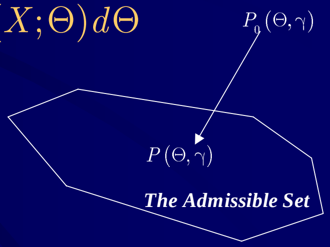
$P_0(\Theta,\gamma)$ = prior over models & margins (favors large margins).

Decision Rule: $\hat{y}=sign\int P(\Theta)L(X;\Theta)d\Theta$

Information Transfer / Projection:

*Information transferred to prior after observations $= KL(P\|P_0)$

*Entropic Regularization and Margin Penalties are on the Same Scale



Maximum Entropy Discrimination Solution

Analytic, Unique, Sparse, Parametric & Structural Models:

$$P(\Theta,\gamma)=\frac{1}{Z(\lambda)}P_0(\Theta,\gamma)\exp\left(\sum_t \lambda_t\left[y_t L(X_t;\Theta)-\gamma_t\right]\right)$$

*$Z(\lambda)$ = normalization constant (partition function)
*$\lambda=\{\lambda_1,\ldots,\lambda_T\}$ =non-negative Lagrange multipliers
* λ solved via unique max of concave objective function:

$$J(\lambda)=-\log Z(\lambda)$$

Example: SVM

$$J(\lambda)=\sum_t\left[\lambda_t+\log\left(1-{}^{\lambda_t}\!/_{c}\right)\right]-\tfrac{1}{2}\sum_{t,t'}\lambda_t\lambda_{t'}y_t y_{t'}\left(X_t^T X_{t'}\right)$$

Example: Generative Models (e-family)

$$L(X;\Theta)=\log\frac{P(X|\theta_+)}{P(X|\theta_-)}+b$$

Use conjugates of $P(X|\theta_\pm)$ for prior
θ = model parameters and structure
b = bias term

Maximum Entropy Discrimination for Regression

Find $P(\Theta,\gamma)$ that mins $KL(P\|P_0)$ subject to constraints:

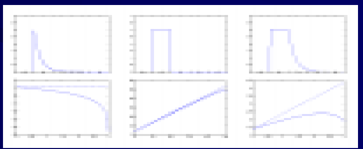
$$\int P(\Theta,\gamma)\left[y_t-L(X_t;\Theta)+\gamma_t\right]d\Theta\, d\gamma \geq 0,\ \forall t$$
$$\int P(\Theta,\gamma)\left[\gamma_t'-y_t+L(X_t;\Theta)\right]d\Theta\, d\gamma \geq 0,\ \forall t$$

Decision Rule: $\hat{y}=\int P(\Theta)L(X;\Theta)d\Theta$

Solution: $P(\Theta,\gamma)=\frac{1}{Z(\lambda)}P_0(\Theta,\gamma)\frac{\exp\left(\sum_t\lambda_t\left[y_t-L(X_t;\Theta)-\gamma_t\right]\right)}{\exp\left(\sum_t\lambda_t'\left[y_t-L(X_t;\Theta)-\gamma_t'\right]\right)}$

Margin Priors: (epsilon-tube)

$$P_0(\gamma_t)\propto\begin{Bmatrix}1 & if 0\leq\gamma_t\leq\in\\ e^{c(\in-\gamma_t)} & if\ \gamma_t>\in\end{Bmatrix}$$



Margin prior

Penalty Fn.

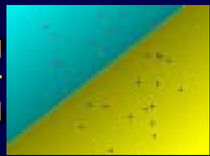
Example: SVM $J(\lambda)=\sum_t y_t(\lambda_t'-\lambda_t)-\in\sum_t(\lambda_t'+\lambda_t)+\sum_t\log(\lambda_t)-\log\left(1-e^{-\lambda_t\in}+\frac{\lambda_t}{c-\lambda_t}\right)$

$$+\sum_t\log(\lambda_t')-\log\left(1-e^{-\lambda_t'\in}+\frac{\lambda_t'}{c-\lambda_t'}\right)-\tfrac{1}{2}\sum_{t,t'}(\lambda_t-\lambda_t')(\lambda_{t'}-\lambda_{t'}')\left(X_t^T X_{t'}\right)$$

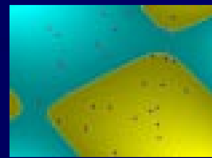
Classification and Regression Examples

Generative Model Classification:

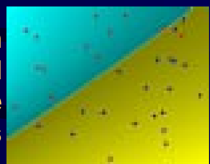
SVM Linear Kernel



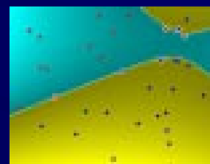
SVM 3rd Order Polynomial



Maximum Likelihood Full Covariance Gaussians

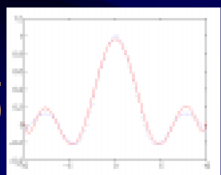


MED Full Covariance Gaussians

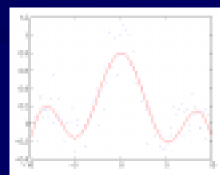


SVM Regression:

Sinc Fn. (Clean)

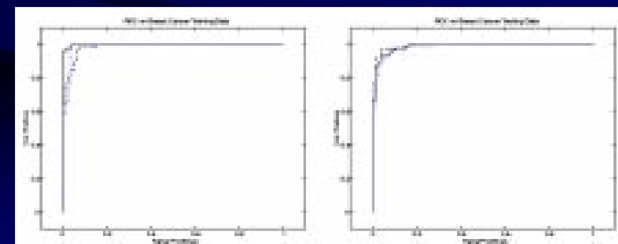


Sinc Fn. With Gaussian Noise



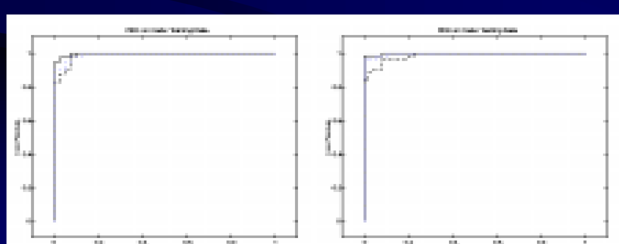
Maximum Entropy Discrimination - Cancer

Method	Training Errors	Testing Errors
Nearest Neighbour		11
SVM - Linear	**8**	**10**
SVM - RBF $\sigma = 0.3$	**0**	**11**
SVM - 3rd Order Polynomial	**1**	**13**
Maximum Likelihood Gaussians	**10**	**16**
MaxEnt Discrimination Gaussians	**3**	**8**



Maximum Entropy Discrimination - Crabs

Method	Training Errors	Testing Errors
Neural Network (1)		3
Neural Network (2)		3
Linear Discriminant		8
Logistic Regression		4
MARS (degree = 1)		4
PP (4 ridge functions)		6
Gaussian Process (HMC)		3
Gaussian Process (MAP)		3
SVM - Linear	**5**	**3**
SVM - RBF $\sigma = 0.3$	**1**	**18**
SVM - 3rd Order Polynomial	**3**	**6**
Maximum Likelihood Gaussians	**4**	**7**
MaxEnt Discrimination Gaussians	**2**	**3**



Feature Selection (Extension)

*Isolates interesting dimensions of the data for the given task
*Typically needs exponential search:
consider all possible subsets of dimensions
*Reduces complexity of data
*Also *Improves Generalization*
*Augments Sparse Vectors (SVMs) with Sparse Dimensions
*Is possible jointly with parameter estimation.
*Can be done discriminatively and efficiently with MED.

Feature Selection

Modify parameters to include a binary ON / OFF ***Switch***

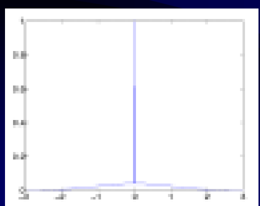
$$L(X;\Theta) = \sum_{i=1}^{n} s_i \theta_i X_i + \theta_0$$

The model $\Theta = \{\theta_0, \dots, \theta_n, s_1, \dots, s_n\}$ contains structural parameters $s_i \in \{0,1\}$ to aggressively prune features.

Prior: $P_0(\Theta) = P_{0,\theta_0}(\theta_0) P_{0,\theta}(\theta) P_{0,s}(s) = P_{0,\theta_0}(\theta_0) N(\theta \,|\, 0, I) \prod_i P_0(s_i)$

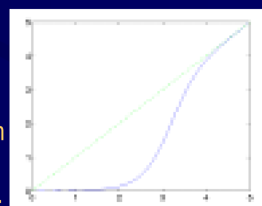
Switch Prior: Bernoulli distribution $P_{s,0}(s_i) = p_0^{s_i}(1-p_0)^{1-s_i}$

p0 parameter smoothly selects no pruning to aggressive pruning



Prior on $s_i\theta_i$

Aggressive attenuation of linear coefficients at low values (p0=.01).

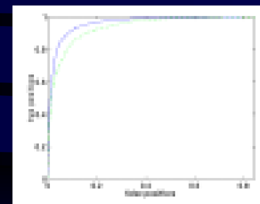


Feature Selection in SVM Classification & Results

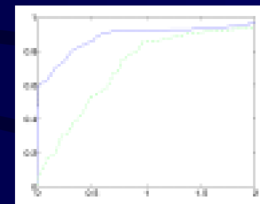
$$J(\lambda) = \sum_t \left[\lambda_t + \log\left(1 - \lambda_t/c\right)\right] - \sum_{i=1}^{n} \log\left[1 - p_0 + p_0 e^{\frac{1}{2}\left(\sum_t \lambda_t y_t X_{t,i}\right)^2}\right]$$

λ constrained to [0,c] hyper-cube with constraint $\sum_t \lambda_t y_t = 0$

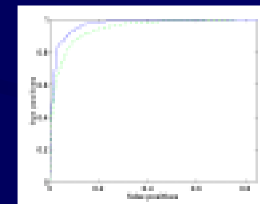
DNA Data: 2-class, 100 element binary vectors. Training Set 500, Testing 4724



ROC of DNA Splice Site
100 Features
Original 25xGATC



CDF of Linear Coeffs
DNA Splice Site
100 Features



ROC DNA Splice Site
~5000 Features
Quadratic Kernel

Dashed line: p0 = 0.99999
Solid line: p0 = 0.00001

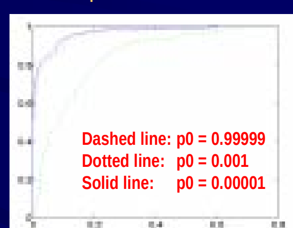
Feature Selection in SVM Regression & Results

$$J(\lambda) = \sum_t y_t(\lambda_t' - \lambda_t) - \epsilon \sum_t (\lambda_t' + \lambda_t) - \tfrac{1}{2}\sigma\left(\sum_t \lambda_t - \lambda_t'\right)^2 + \sum_t \log(\lambda_t) - \log\left(1 - e^{-\lambda_t \epsilon} + \tfrac{\lambda_t}{c-\lambda_t}\right)$$

$$+ \sum_t \log(\lambda_t') - \log\left(1 - e^{-\lambda_t' \epsilon} + \tfrac{\lambda_t'}{c-\lambda_t'}\right) - \sum_i \log\left(1 - p_0 + p_0 e^{\frac{1}{2}\left[\sum_t (\lambda_t - \lambda_t') X_{t,i}\right]^2}\right)$$

Boston Housing Data: 13 scalar features. Training Set 481, Testing 25
Explicit Quadratic Kernel Expansion Used

Linear Model Estimator	Epsilon-Sensitive Linear Loss
Least-Squares	1.7584
MED p0 = 0.99999	1.7529
MED p0 = 0.1	1.6894
MED p0 = 0.001	1.5377
MED p0 = 0.00001	1.4808



D. Ross Cancer Data: 67 scalar features. Training Set 50, Testing 3951

Linear Model Estimator	Epsilon-Sensitive Linear Loss
Least-Squares	3.609e+03
MED p0 = 0.00001	1.6734e+03

Feature Selection in Generative Models

Feature selection is not limited to SVMs.
Applies to discriminative Generative Model Estimation as well.
But, tractable computation sometimes needs approximations.

Example: 2-class Gaussian distributions
variable means, identity covariance

Parameters: $\{\mu, \nu\}$ Prior: $P_0(\mu) \sim P_0(\nu) \sim N(0, I)$
Switches: $\{s, r\}$ Prior: $P_0(s_i) \sim P_0(r_i) = p_0^{r_i}(1-p_0)^{1-r_i}$

Discriminant Function (tractable in this case):

$$L(X;\Theta) = \log \frac{P(X|\theta_+)}{P(X|\theta_-)} + b = \sum_i s_i (X_i - \mu_i)^2 - \sum_i r_i (X_i - \nu_i)^2 + b$$

Latent Transformations (Extension)

Each input example has additional unobservable properties
Only have a prior distribution over the unobservable
I.e. category, affine transform, latent variable, alignment

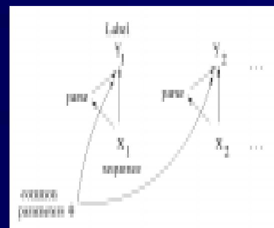
Given: training examples: $\{X_1, \ldots, X_T\}$

binary (+/- 1) labels: $\{y_1, \ldots, y_T\}$

hidden transformations: $\{U_1, \ldots, U_T\}$

transformation function: $\hat{X} = T(X, U)$

prior on transforms: $P_0(U_t)$



Solution: $P(\Theta, U, \gamma) = \frac{1}{Z(\lambda)} P_0(\Theta, U, \gamma) \exp\left(\sum_t \lambda_t \left[y_t L\left(T(X_t, U_t); \Theta\right) - \gamma_t\right]\right)$

Transductive and Iterative. Solve iteratively by alternating solution of P(θ) and P(U).

Example: $L(\hat{X}_t; \Theta) = L\left(T(X_t, U_t); \Theta\right) = \theta^T\left(X_t - U_t \vec{1}\right) + b$

Optimization & Bounded QP (Extension)

MED maximizes concave objective with convex constraints.
Axis-parallel, Newton, gradient descent will converge.

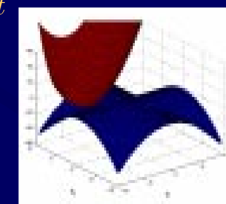
Lower Bound the concave objective with quadratic
Can then use SMO, QP, and other SVM optimizers.

Example: SVM Feature Selection

$$j_i(\lambda) = -\log\left(1 - p_0 + p_0 e^{\frac{1}{2}\left[\sum_t (\lambda_t - \lambda_t') X_{t,i}\right]^2}\right) = -\log\left(1 - p_0 + p_0 e^{\frac{1}{2}\vec{\lambda}^T M \vec{\lambda}}\right)$$

$$j_i(\lambda) \geq \vec{\lambda}^T (N + hM)\tilde{\lambda} - \tfrac{1}{2}\vec{\lambda}^T (M + N)\vec{\lambda} + const$$

$$where\, N = \tfrac{1}{4}\left(M\tilde{\lambda}\right)\left(M\tilde{\lambda}\right)^T \; and\, h = \frac{1 - p_0}{1 - p_0 + p_0 \exp\left(\frac{1}{2}\tilde{\lambda}^T M \tilde{\lambda}\right)}$$



Iterate bound (contact at $\tilde{\lambda}$) and QP
Each QP is seeded at previous sol'n
Converges in about 10 fast iterations

MED for the Exponential Family (Extension)

Proof: MED with generative models spans members of the exponential family (where Gaussians generate SVMs):

exponential family form: $p(X \mid \theta) = \exp\left(A(X) + X^T\theta - K(\theta)\right)$

conjugate prior: $p(\theta \mid \chi) = \exp\left(\tilde{A}(\theta) + \theta^T\chi - \tilde{K}(\chi)\right)$

Analytic Partition Function for Classification:

Using Non-Informative Prior on b

$$Z_\Theta = \int P_0(\Theta) \exp\left(\sum_t \lambda_t y_t L(X_t; \Theta)\right) d\Theta$$

$$Z_\Theta = \int P_0(\theta_+) P_0(\theta_-) P_0(b) \exp\left(\sum_t \lambda_t y_t \left[\log \frac{P(X|\theta_+)}{P(X|\theta_-)} + b\right]\right) d\Theta$$

$$Z_{\theta_\pm} = \int \exp\left(\tilde{A}(\theta_\pm) + \theta_\pm^T \chi - \tilde{K}(\chi)\right) \exp\left(\sum_t \lambda_t y_t \left(A(X_t) + X_t^T \theta_\pm - K(\theta_\pm)\right)\right) d\theta_\pm$$

$$Z_{\theta_\pm} = \exp\left(-\tilde{K}(\chi) + \sum_t \lambda_t y_t A(X_t)\right) \times \int \exp\left(\tilde{A}(\theta_\pm) + \theta_\pm^T \left(\chi + \sum_t \lambda_t y_t X_t\right)\right) d\theta_\pm$$

$$Z_{\theta_\pm} = \exp\left(-\tilde{K}(\chi) + \sum_t \lambda_t y_t A(X_t)\right) \times \exp\left(\tilde{K}\left(\chi + \sum_t \lambda_t y_t X_t\right)\right)$$

Concluding Ideas on MED and Feature Selection

Maximum Entropy Discrimination is a flexible Bayesian regularization approach. It provides a geometric view of learning as constrained minimization to prior distributions over margins, parameters, latent variables. It simultaneously combines:

- probabilistic methods
- large margin discrimination and SVMs
- feature selection
- classification, regression, etc.
- parameter and structure estimation
- exponential family generative models
- transduction and detection

Feature Selection is a particularly advantageous extension which provides increased sparsity (support vectors & support dimensions) and improves generalization.

Limitation of MED

Applies to Exponential Family
Yet many models are MIXTURES of E-family:

Latent Models
HMMs
Mixture Models
Incomplete Data
Hidden Variable Bayesian Networks

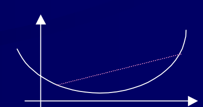
Intractable models in discriminative & conditional settings
Thus use Variational Bounds to Perform Calculations
Invoke EM and derive its Discriminative DUAL
by Reversing Jensen

Reversing Jensen's Inequality

The Dual of EM for Discriminative Latent Learning

Tony Jebara
Alex Pentland

Jensen's Inequality



-Inequalities allow us to Integrate, Maximize,
Evaluate and Derive Intractable Expressions
-Convexity: 1905-1906 by J. Jensen (Dutch Mathematician & Engineer)
-See "Convex Functions, Partial Ordering and Statistical Applications"
by J. Pecaric, F. Proschan and Y. Tong.

Jensen in Statistics and EM:
-Subsumes many information theoretic bounds (Cover & Thomas)
-Subsumes the EM Algorithm (Demspter, Laird & Rubin, Baum-Welch)
-EM casts latent variable problems as complete data by solving
for a lower bound on likelihood.

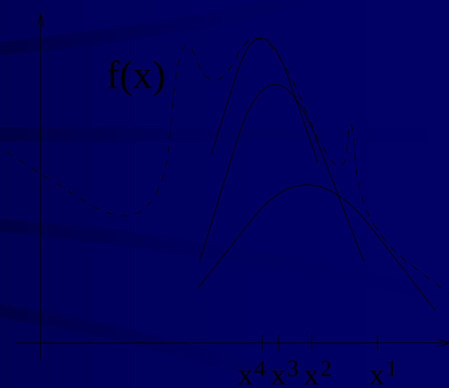
Reversals of Jensen:
-Constrained reversals and converses have been explored and are active
areas in mathematics (S.S. Dragomir).
-Reversals have yet to be applied to ***discriminative learning***.

The EM Algorithm

Makes intractable maximization of likelihood and integration of Bayesian inference tractable via variational bounds.

E-step: Replace unknowns with their expected values under current model.
(i.e. solving for a lower bound on likelihood using Jensen!)

M-step: Optimize current model with the complete data
(maximizing the Jensen lower bound!)



Applies and converges for Exponential Family Mixtures. I.e. a very large space of models that covers most of contemporary machine learning. HMMs, Gaussian Mixture Models, etc.

The Exponential Family

E-family: K() is a convex function, X in its gradient space

Special Properties: conjugates, linearity, convexity, etc. TRACTABLE

$$P(X \mid \Theta) = \exp\left(A(X) + X^T\theta - K(\theta)\right)$$

Includes Poisson, Gaussian, Trees, Multinomial, Exponential

E-Distribution	$A(X)$	$K(\theta)$
Gaussian	$-\frac{1}{2}X^TX - \frac{D}{2}\log(2\pi)$	$-\frac{1}{2}\theta^T\theta$
Multinomial	0	$\eta\log\left(1+\sum_d \exp\theta_d\right)$
Exponential	0	$-\log(-\theta) \quad \forall\,\theta < 0$

Mixtures of E-family: Gaussian mixture models, HMMs, Sigmoidal Belief Nets, Latent Bayes Nets,etc. Most machine learning models!

INTRACTABLE

$$P(X \mid \Theta) = \sum_m p(m)\, p(X \mid m, \theta)$$

$$= \sum_m \alpha_m \exp\left(A_m(X_m) + X_m{}^T\theta_m - K_m(\theta_m)\right)$$

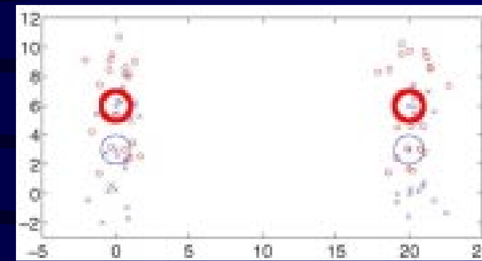
$$= \sum_m \sum_n \alpha_{mn} \exp\left(A_{mn}(X_{mn}) + X_{mn}{}^T\theta_m - K_m(\theta_m)\right)$$

Conditional & Discriminative Classification

TWO WHITE GAUSSIANS PER CLASS FOR MODEL.

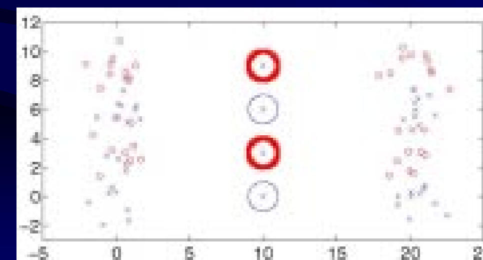
DATA REALLY COMES FROM FOUR WHITE GAUSSIANS

EM



L = - 8.0
L_c = - 1.7

CEM

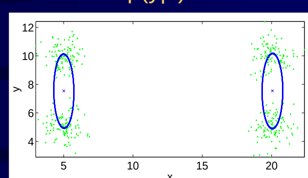


L = - 54.7
L_c = + 0.4

Conditional & Discriminative Regression

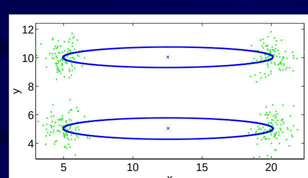
p(y|x)

PROBLEMATIC MAXIMUM LIKELIHOOD SITUATIONS

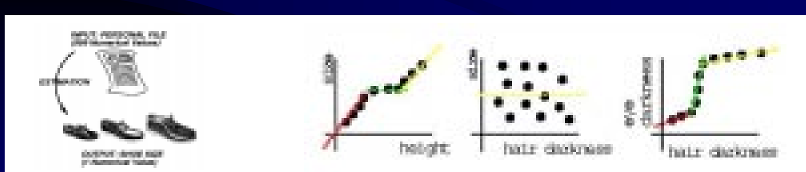


L = - 4.2
L_c = - 2.4

CONDITIONAL MAXIMUM LIKELIHOOD SOLUTIONS



L = - 5.2
L_c = - 1.8



Discriminative Criteria and Negated Log-Sums

Maximum Likelihood: Clusters towards all data

$$l = \sum_i \log \sum_m p(m, c_i, X_i \mid \theta)$$

Maximum Conditional Likelihood: Emphasizes classification task

$$l^c = \sum_i \log \sum_m p(m, c_i \mid X_i, \theta)$$

Repels models away from incorrect data

$$= \sum_i \log \sum_m p(m, c_i, X_i \mid \theta) - \sum_i \log \sum_m \sum_c p(m, c, X_i \mid \theta)$$

Maximum Margin Discrimination (MED): Emphasizes sparsity and discriminant boundary for task

$$L(X \mid \theta) = \log \frac{p(X \mid \theta_+)}{p(X \mid \theta_-)}$$

$$= \log \sum_m p(m, X \mid \theta_+) - \log \sum_m p(m, X \mid \theta_-)$$

The above log-sums make integration, maximization, etc. intractable. Need to simplify via overall lower and upper bounds...

Jensen $f(E\{\square\}) \geq E\{f(\square)\}$

DANGER: IF WE HAVE NEGATED LOG- SUM GET UPPER BOUND INSTEAD!

Uses Concavity of log()
Reweights data with responsibilties
Variational Lower Bound at Current Model (θ)
makes tangential contact with true objective function log-sum
Local Computations to get a Global Lower Bound

$$\log \sum_m p(m, X|\Theta) \;\geq\; \sum_m \left(\frac{p(m,X|\hat{\Theta})}{\sum_n p(n,X|\hat{\Theta})} \right) \log \frac{p(m,X|\Theta)}{p(m,X|\hat{\Theta})} + \log \sum_m p(m,X|\hat{\Theta})$$

$$\log \sum_m \alpha_m \exp(\mathcal{A}_m(X_m) + X_m^T \Theta_m - \mathcal{K}_m(\Theta_m)) \;\geq\; \sum_m -w_m (Y_m^T \Theta_m - \mathcal{K}_m(\Theta_m)) + k$$

$$k \;=\; \log p(X|\hat{\Theta}) + \sum_m w_m (Y_m^T \hat{\Theta}_m - \mathcal{K}_m(\hat{\Theta}_m))$$

$$Y_m \;=\; \frac{1}{w_m} h_m \left(\left.\frac{\partial \mathcal{K}_m(\Theta_m)}{\partial \Theta_m}\right|_{\hat{\Theta}_m} - X_m \right) + \left.\frac{\partial \mathcal{K}_m(\Theta_m)}{\partial \Theta_m}\right|_{\hat{\Theta}_m} \;=\; X_m$$

$$w_m \;=\; -h_m \;=\; -\left(\frac{p(m,X|\hat{\Theta})}{\sum_n p(n,X|\hat{\Theta})} \right)$$

Reverse-Jensen $f(E\{\square\}) \leq E\{f(\square)\}$

Uses convexity of E-family
Reweights and Translates data
Variational Upper Bound at Current Model (θ)
makes tangential contact with true objective function log-sum
Local Computations to get a Global Upper Bound

$$\log \textstyle\sum_m \alpha_m \exp(\mathcal{A}_m(X_m) + X_m^T \Theta_m - \mathcal{K}_m(\Theta_m)) \;\leq\; \textstyle\sum_m -[w_m] (Y_m^T \Theta_m - \mathcal{K}_m(\Theta_m)) + k$$

$$k \;=\; \log p(X|\tilde{\Theta}) + \sum_m w_m (Y_m^T \tilde{\Theta}_m - \mathcal{K}_m(\tilde{\Theta}_m))$$

$$Y_m \;=\; \frac{h_m}{w_m} \left(\left.\frac{\partial \mathcal{K}(\Theta_m)}{\partial \Theta_m}\right|_{\tilde{\Theta}_m} - X_m \right) + \left.\frac{\partial \mathcal{K}(\Theta_m)}{\partial \Theta_m}\right|_{\tilde{\Theta}_m}$$

$$w'_m \quad \#1 \rightarrow \quad \min w'_m \text{ such that } \frac{h_m}{w'_m} \left(\left.\frac{\partial \mathcal{K}(\Theta_m)}{\partial \Theta_m}\right|_{\tilde{\Theta}_m} - X_m \right) + \left.\frac{\partial \mathcal{K}(\Theta_m)}{\partial \Theta_m}\right|_{\tilde{\Theta}_m} \in \frac{\partial \mathcal{K}(\Theta_m)}{\partial \Theta_m}$$

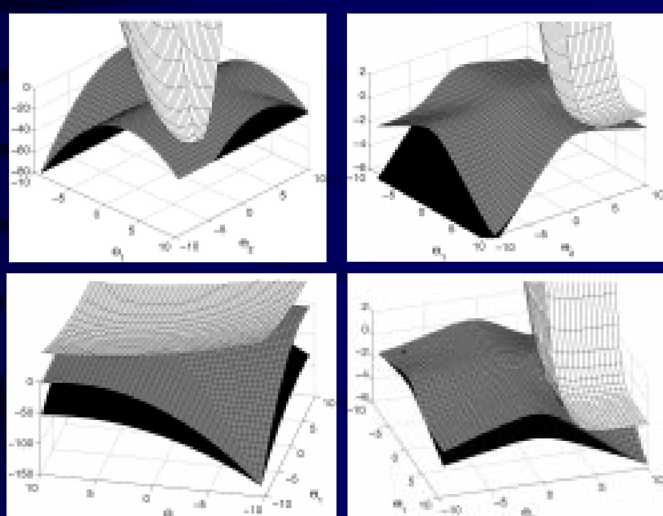
$$w_m \quad \#2 \rightarrow \quad w_m \;\geq\; \frac{1}{2} \left[X_m - \mathcal{K}'(\tilde{\Theta}_m) \right]^T \mathcal{K}''(\tilde{\Theta}_m)^{-1} \left[X_m - \mathcal{K}'(\tilde{\Theta}_m) \right] + w'_m$$

OR (tighter...)

$$w_m \quad \#2 \rightarrow \quad w_m \;\geq\; \frac{1}{-2\log(h_m)} \left[X_m - \mathcal{K}'(\tilde{\Theta}_m) \right]^T \mathcal{K}''(\tilde{\Theta}_m)^{-1} \left[X_m - \mathcal{K}'(\tilde{\Theta}_m) \right] + w'_m$$

Reverse-Jensen Bounds

Log-sum (gray)
Jensen (black)
Rev-Jensen (white)



Gaussian Case | Multinomial Case

$$\sum_m w_m \left(\mathcal{K}(\Theta_m) - \mathcal{K}(\tilde{\Theta}_m) - (\Theta_m - \tilde{\Theta}_m)^T \mathcal{K}'(\tilde{\Theta}_m) \right) \;\geq\; \log \frac{p(X|\Theta)}{p(X|\tilde{\Theta})} + \sum_m h_m (\Theta_m - \tilde{\Theta}_m)^T \left(\mathcal{K}'(\tilde{\Theta}_m) - X_m \right)$$

Define $\mathcal{F}_m(\Theta_m) = \mathcal{K}(\Theta_m) - \mathcal{K}(\tilde{\Theta}_m) - (\Theta_m - \tilde{\Theta}_m)^T \mathcal{K}'(\tilde{\Theta}_m)$ and $Z_m = X_m - \mathcal{K}'(\tilde{\Theta}_m)$.

$\sum_m w_m \mathcal{F}(\Theta_m) \;\geq\; \log \frac{\sum_m \exp\{D_m + \Theta_m^T Z_m - \mathcal{F}(\Theta_m)\}}{\sum_m \exp\{D_m + \tilde{\Theta}_m^T Z_m - \mathcal{F}(\tilde{\Theta}_m)\}} - \sum_m h_m (\Theta_m - \tilde{\Theta}_m)^T Z_m$

Define: $\mathcal{F}_m(\Theta_m) \;=\; \mathcal{G}_m(\Phi_m) \;=\; \frac{1}{2}(\Phi_m - \tilde{\Theta}_m)^T (\Phi_m - \tilde{\Theta}_m)$

$$\sum_m w_m \mathcal{G}(\Phi_m) \;\geq\; \log \frac{\sum_m \exp\{D_m + \Theta_m(\Phi_m)^T Z_m - \mathcal{G}(\Phi_m)\}}{\sum_m \exp\left\{D_m + \tilde{\Theta}_m^T Z_m - \mathcal{G}(\tilde{\Theta}_m)\right\}} - \sum_m h_m (\Theta_m(\Phi_m) - \tilde{\Theta}_m)^T Z_m \quad (1)$$

Hessian of $\mathcal{F} = \mathcal{G}$: $\mathcal{K}''(\Theta_m) \frac{\partial \Theta_m}{\partial \Phi_m} \frac{\partial \Theta_m}{\partial \Phi_m}^T + (\mathcal{K}'(\Theta_m) - \mathcal{K}'(\tilde{\Theta}_m)) \frac{\partial^2 \Theta_m}{\partial \Phi_m^2} \;=\; I$

At $\Theta_m = \tilde{\Theta}_m$: $\left.\frac{\partial \Theta_m}{\partial \Phi_m}\right|_{\tilde{\Theta}_m} \;=\; [\mathcal{K}''(\tilde{\Theta}_m)]^{-1/2}$.

In an e-family, we can always find a Θ_m^* such that $X_m = \mathcal{K}'(\Theta_m^*)$. By convexity of $\mathcal{F}$ we create a linear lower bound at Θ_m^*:

$$\mathcal{F}(\Theta_m^*) + (\Theta_m - \Theta_m^*) \left.\frac{\partial \mathcal{F}(\Theta_m)}{\partial \Theta_m}\right|_{\Theta_m^*} \;\leq\; \mathcal{F}(\Theta_m) \;=\; \mathcal{G}(\Phi_m)$$

Short Proof

Taking 2nd derivatives in Φ_m gives: $\mathcal{F}'(\Theta_m^*) \frac{\partial^2 \Theta_m}{\partial \Phi_m^2} \;\leq\; I$ or $Z_m \frac{\partial^2 \Theta_m}{\partial \Phi_m^2} \;\leq\; I$

Thus, $D_m + \Theta_m(\Phi_m)^T Z_m - \mathcal{G}(\Phi_m)$ is concave, linear upper bound at $\tilde{\Theta}_m$:

$$\sum_m w_m \mathcal{G}(\Phi_m) \;\geq\; \log \frac{\sum_m \exp\left\{D'_m + \Phi_m^T [\mathcal{K}''(\tilde{\Theta}_m)]^{-1/2} Z_m\right\}}{\sum_m \exp\left\{D_m + \tilde{\Theta}_m^T Z_m - \mathcal{G}(\tilde{\Theta}_m)\right\}} - \sum_m h_m (\Theta_m(\Phi_m) - \tilde{\Theta}_m)^T Z_m$$

map bowl to bowl

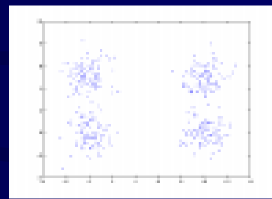
Taking 2nd derivatives over Φ_m: $w_m I \;\geq\; \frac{1}{2} Z_m \mathcal{K}''(\tilde{\Theta}_m)^{-1} Z_m^T \;-\; h_m Z_m \frac{\partial^2 \Theta_m}{\partial \Phi_m^2}$

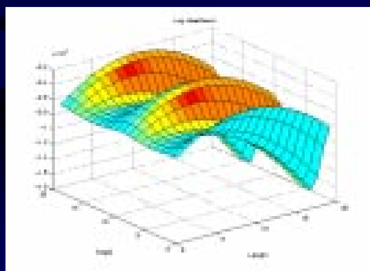
Invoke constraint #1 and replace $-h_m Z_m \frac{\partial^2 \Theta_m}{\partial \Phi_m^2} \;\leq\; w'_m I$. Manipulate to obtain:

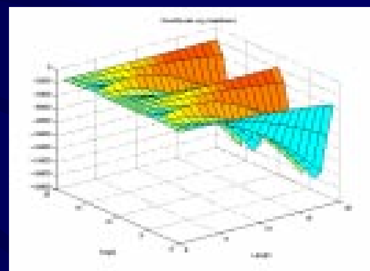
$$w_m I \;\geq\; \frac{1}{2}[X_m - \mathcal{K}'(\tilde{\Theta}_m)] \mathcal{K}''(\tilde{\Theta}_m)^{-1} [X_m - \mathcal{K}'(\tilde{\Theta}_m)]^T + w'_m I \qquad \square$$

CEM Regression Results

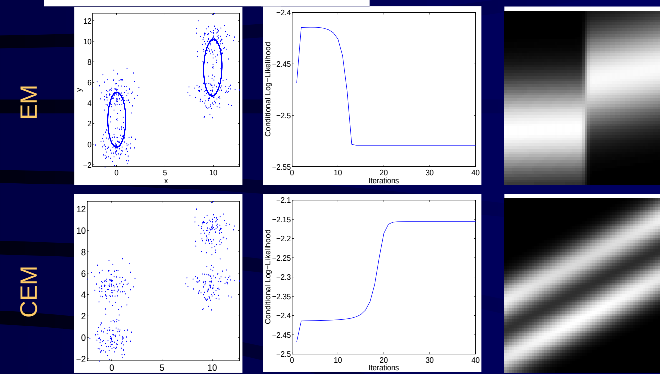
Estimate p(y|x) regression model with 2 Gaussians (Gaussian gates with linear experts).







CEM Regression Results



CEM monotonically increases conditional likelihood unlike EM. Result: better p(y|x) which captures the multimodality in y without wasting resources in x.

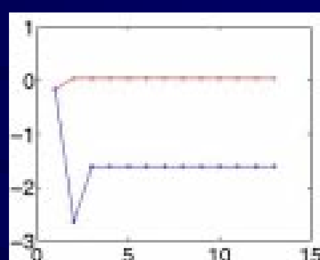
Abalone age data set (UCI).

Algorithm	Regression Accuracy
Cascade-Correlation (0 hidden)	24.86%
Cascade-Correlation (5 hidden)	26.25%
C4.5	21.5%
Linear Discriminant	0.0%
k=5 Nearest Neighbor	3.57%
EM 2 Gaussians	22.32%
EM&CEM 1 Gaussian	20.79%
CEM 2 Gaussians	**26.63%**

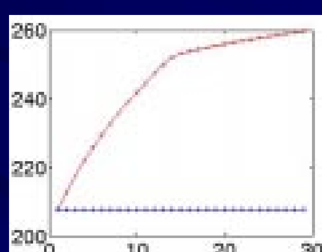
CEM Classification Results

Gaussian mixture model shown earlier for classification. Monotonic convergence. Double computation of EM per epoch.

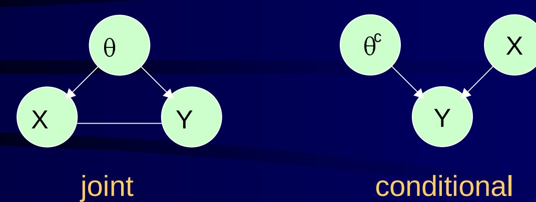
CEM accuracy = 93%
EM accuracy = 59%



Multinomial mixture model. 3-class multinomials for 60 base-pair protein chains. CEM monotonically increases conditional likelihood.



Bayesian Inference: Conditional vs. Joint



* Each assumes different independencies in data

= Data Structure (like Model Structure which is useful for learning)

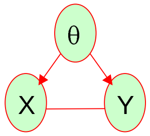
* Exponential number of conditional models

-> Use a handful for frequent task and joint for rare task

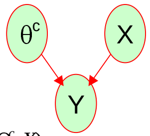
-> Use marginal for unreliable covariates

Bayesian Approach

$$p(x,y) = p(x,y|X,Y)$$
$$= \int p(x,y,\Theta|X,Y)\,d\Theta$$
$$= \int p(x,y|\Theta)\,p(\Theta|X,Y)\,d\Theta$$
$$p(y|x)^j = \frac{\int p(x,y|\Theta)\,p(\Theta|X,Y)\,d\Theta}{\int p(x|\Theta)\,p(\Theta|X,Y)\,d\Theta}$$
$$p(x,y|X,Y) \approx p(x,y|\Theta^*)$$
where
$$\Theta^* = \begin{cases} MAP \rightarrow \arg\max p(\Theta|X,Y) = \arg\max p(X,Y|\Theta)\,p(\Theta) \\ ML \rightarrow \qquad \arg\max p(X,Y|\Theta) \end{cases}$$

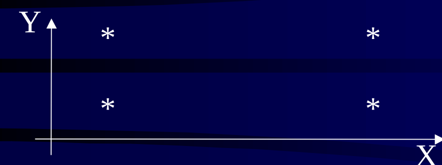


$$p(y|x)^c = p(y|x,X,Y)$$
$$= \int p(y,\Theta^c|x,X,Y)\,d\Theta^c$$
$$= \int p(y|x,\Theta^c)\,p(\Theta^c,X,Y)\,d\Theta^c$$
$$p(\Theta^c|X,Y) = \frac{p(Y|\Theta^c,X)\,p(\Theta^c,X)}{p(X,Y)}$$
$$= \frac{p(Y|\Theta^c,X)\,p(X|\Theta^c)\,p(\Theta^c)}{p(X,Y)}$$
$$= \frac{p(Y|\Theta^c,X)\,p(X)\,p(\Theta^c)}{p(X,Y)}$$
$$p(y|x)^c = \int p(y|x,\Theta^c)\frac{p(Y|\Theta^c,X)\,p(X)\,p(\Theta^c)}{p(X,Y)}\,d\Theta^c$$
$$= \int p(y|x,\Theta^c)\,p(Y|\Theta^c,X)\,p(\Theta^c)\,d\Theta^c \Big/ p(Y|X)$$
$$p(y|x) \approx p(y|x,\Theta^{*c})$$
$$\textit{where}\ \Theta^{*c} = \begin{cases} MAP \rightarrow \arg\max p(Y|\Theta^c,X)\,p(\Theta^c) \\ ML \rightarrow \ \arg\max p(y|\Theta^c,X) \end{cases}$$

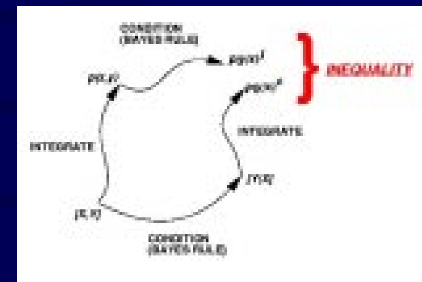


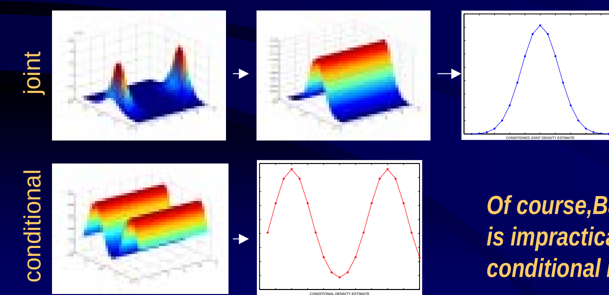
x: covariate
y: response
X: covariate dataset
Y: response dataset

An Exact Bayesian Example



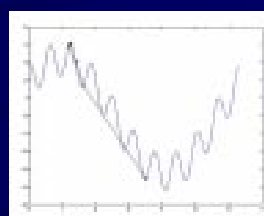
Data: p(y|x) from 4 points to fit with 2 Gaussians





Of course,Bayesian integration is impractical so we must consider conditional MAP and ML...

Extensions



Annealing
Bound can accommodate a Gibbs temperature for global optimum.

$$\exp\left(\beta \times \left[A(X) + X^T\theta - K(\theta)\right]\right)$$

Latent Bayesian Networks

Hidden Markov Models

MED - Large Margin Latent Discrimination

Variational Bayesian Inference

Generic Optimization

check http://www.media.mit.edu/~jebara

Action-Reaction Learning

Automatic Visual Analysis and Synthesis of Interactive Behaviour

Tony Jebara
Alex Pentland

Motivation & Background

IDEA
Computer Vision, Face & Gesture at Killington, Behaviourists...

BEHAVIOUR PERCEPTION
Static Imagery -> Simple Temporal Models -> Learned Temporal Dynamics, Higher Order Control, Multiple Hypothesis, HMMs, NNs (Blake, Bregler, Pentland, Bobick, Hogg)

BEHAVIOUR SYNTHESIS
Competing Behaviours, Control, Reinforcement, Ethology, Cog-Sci (Brooks, Terzopolous, Blumberg, Uchibe, Mataric, Large)

ARL (ACTION REACTION LEARNING)
-Machine Learning of Correlations between Stimulus & Response via Perceptual Measurements of Human Interactions
-Imitation Based Learning (Mataric)
-Behaviourism (Thorndike, Watson, Skinnerian, Gibsonian) -> Reactionary
-Watch Humans Interacting to Learn how to React to Stimulus

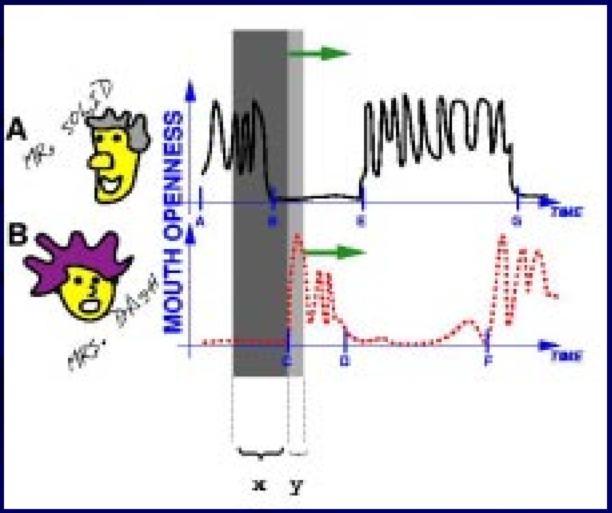
Scenario

AUTOMATIC UNSUPERVISED OBSERVATION OF 2 AGENT INTERACTION

TRACK LIP MOTIONS

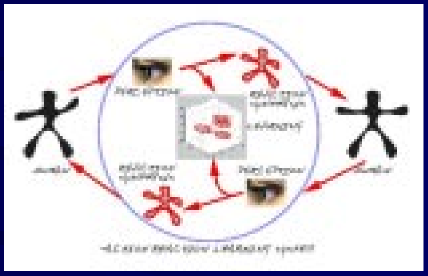
DISCOVER CORRELATIONS BETWEEN PAST ACTION & CONSEQUENT REACTION

ESTIMATE p(y|x)

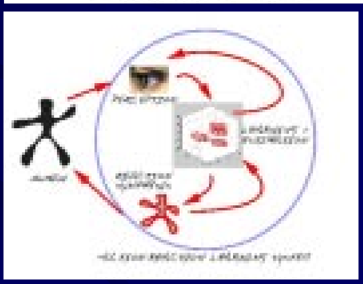


System Architecture

OFFLINE: LEARNING FROM HUMAN INTERACTION, SPYING ON TWO USERS TO LEARN p(y|x)



ONLINE: INTERACTION WITH SINGLE USER WITH LEARNED p(y|x)



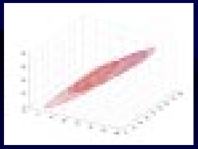
Perception

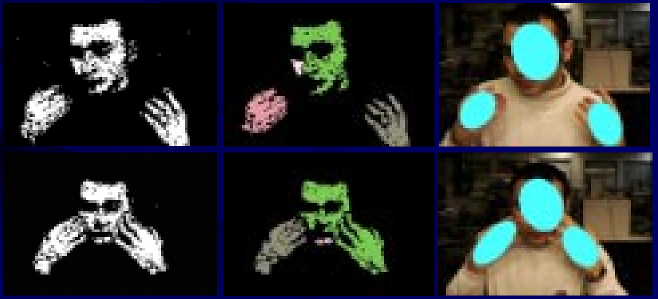
PROBABILISTIC HEAD & HAND TRACKING

EXPECTATION MAXIMIZATION (EM)

$$p(\mathbf{x}_{rgb}) = \sum_{i=1}^{3} \frac{p(i)}{(2\pi)^{3/2}\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x}_{rgb}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}_{rgb}-\mu_i)}$$

$$p(\mathbf{x}_{xy}) = \sum_{j=1}^{3} \frac{p(j)}{2\pi\sqrt{|\Sigma_j|}} e^{-\frac{1}{2}(\mathbf{x}_{xy}-\mu_j)^T \Sigma_j^{-1}(\mathbf{x}_{xy}-\mu_j)}$$



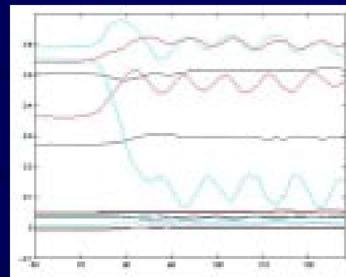


Perception...

TEMPORAL REPRESENTATION

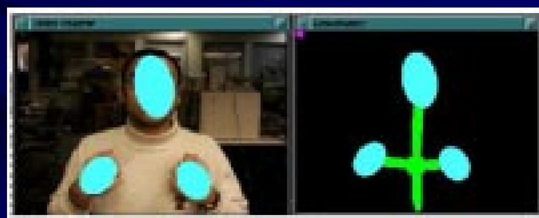
5 Parameters per Blob =
2 Centroid +
3 Square Root Covariance

$$\mu_x \mu_y \Gamma_{xx} \Gamma_{xy} \Gamma_{yy}$$



GRAPHICAL OUTPUT

(Seen by both users)

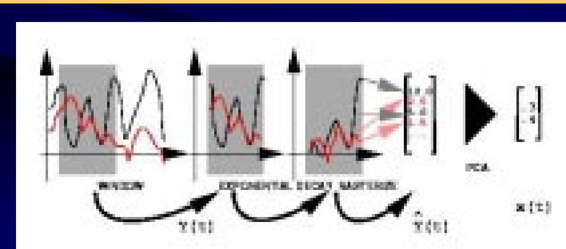


Temporal Modeling

TIME SERIES PREDICTION (Gershenfeld, Weigund, Mozer, Wan)
Santa Fe: NNs, RNNs, HMMs, Diff Eqns, etc.
Sun Spot, Bach, Physiological, Chaotic Laser

SHORT TERM MEMORY PRE-PROCESSING (Wan, Elliot-Anderson)

$\mathbf{y} = [B_{a1} B_{a2} B_{a3} B_{b1} B_{b2} B_{b3}]$	Features (Blobs Concatenated)
$\mathbf{y}(t) \approx g(\mathbf{y}(t-1), \mathbf{y}(t-2), \ldots, \mathbf{y}(t-T))$	Prediction Mapping
$Y(t) = [\mathbf{y}(t-1)\mathbf{y}(t-2) \ldots \mathbf{y}(t-T)]$	Short Term Memory (T=120)



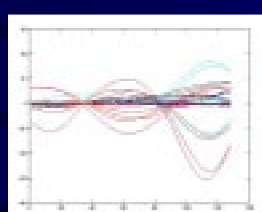
Temporal Modeling...

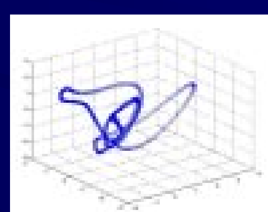
PRINCIPAL COMPONENTS ANALYSIS

(linear, FFT, Wavelets...)

Gaussian Distribution of STM (roughly 6.5 seconds)
Dims = T x Feats = 120x30 ---> 40 (95% Energy of Submanifold)
Low dimensional (i.e. smoothed) characterization of past interaction


eigenvalues


eigenvectors


eigenspace

LEARN MAPPING PROBABILISTICALLY

p(y|x) = p(future | STM) versus deterministic y=g(x)

Learning & The CEM Algorithm

EXPECTATION MAXIMIZATION (EM)

Learns p(x,y) by maximizing $\prod_{i=1}^{N} p(x_i, y_i \mid \Theta)$ (joint model of phenomenon)

Powerful convergence -- Clean statistical Framework
More global than gradient solutions -- Can be deterministically annealed

For conditional problems (input/output regression, classification)
Joint models are outperformed (I.e. NNs and RNNs versus HMMs)
Since they don't optimize output error (I.e. testing criterion is not like training)

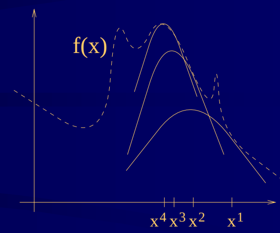
CONDITIONAL EXPECTATION MAXIMIZATION (CEM) *NIPS11,1998*

Learns p(y|x) by maximizing $\prod_{i=1}^{N} p(y_i \mid x_i, \Theta)$ (conditional model of task)

Convergence properties like EM but for Conditional Likelihood

Learning and the CEM Algorithm...

Variational Bound Maximization
- Monotonic
- Skips Local Minima
- Deterministically Annealable



Model: Mixture of Experts
- Soft Piece-Wise Linear
- Gaussian Gates with Conditioned Gaussian Regressors
- CEM Applies to other models: HMM, Multinomial Mixtures, etc.

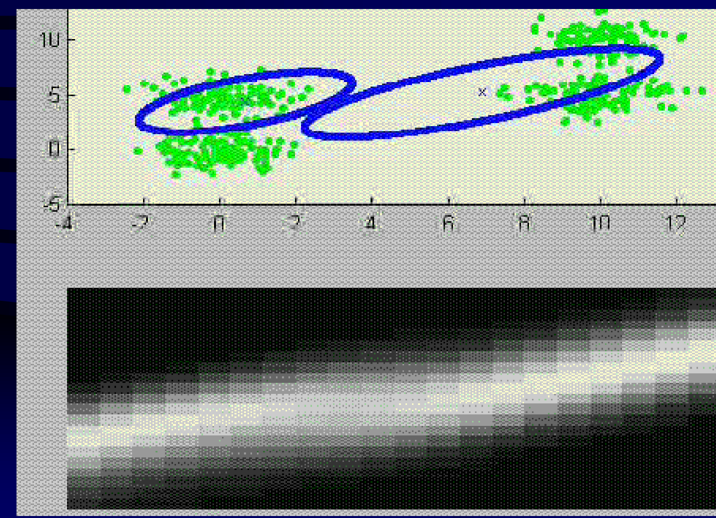
$$p(\mathbf{y}\mid\mathbf{x},\Theta)=\frac{\sum_{m=1}^{M}\alpha_m \bar{N}(\mathbf{x};\mu_m,\Sigma_m)\times N(\mathbf{y};\upsilon_m+\Gamma_m\mathbf{x},\Omega_m)}{\sum_{m=1}^{M}\alpha_m \bar{N}(\mathbf{x};\mu_m,\Sigma_m)}$$

Output: Expectation or Arg Max (Regression)

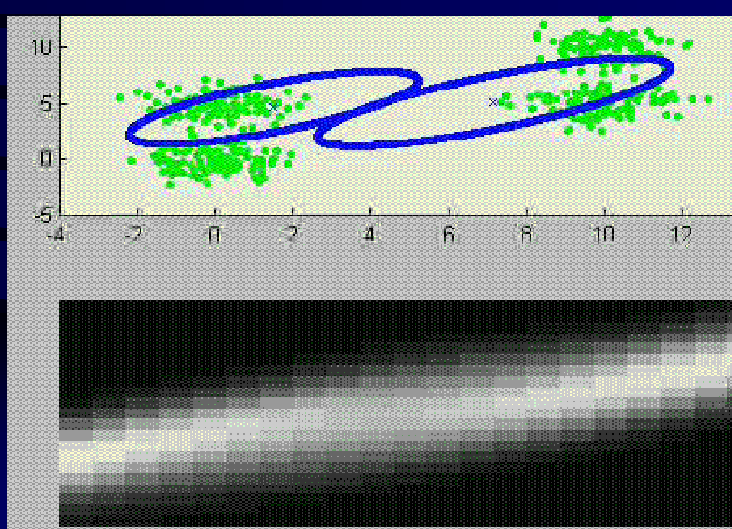
$$\hat{\mathbf{y}}=\int \mathbf{y}\,p(\mathbf{y}\mid\hat{\mathbf{x}})d\mathbf{y}=\frac{\sum_{m=1}^{M}\hat{\mathbf{y}}_m\, p(\hat{\mathbf{y}}_m\mid\hat{\mathbf{x}})}{\sum_{m=1}^{M}p(\hat{\mathbf{y}}_m\mid\hat{\mathbf{x}})} \qquad \hat{\mathbf{y}}_m=\mu_m^y+\Sigma_m^{yx}\Sigma_m^{xx^{-1}}(\hat{\mathbf{x}}-\mu_m^x)$$

Computationally very efficient, simple matrix operations

EM



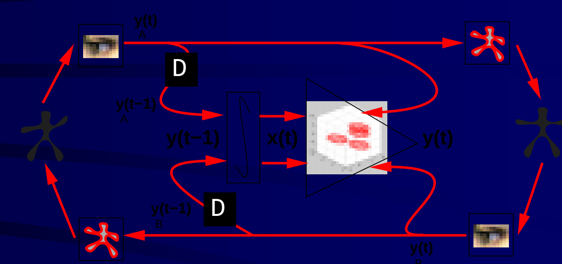
CEM



Integration

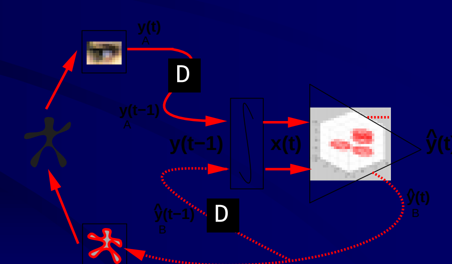
TRAINING MODE

System accumulates action / reaction pairs (x,y) and uses CEM to optimize a conditioned Gaussian mixture model for p(y|x)



INTERACTIVE MODE

System completes missing time series and synthesizes reaction in graphical form for user using the p(y|x)



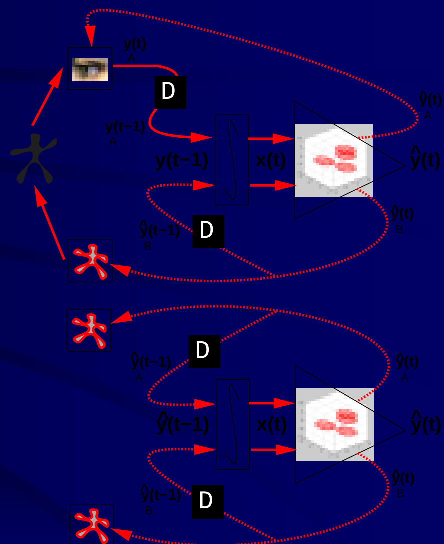
Integration...

FEEDBACK MODE

Predicted measurement on user can be fed back as a non-linear learned EKF to aid vision. Can also use p(x) to filter vision and find correspondence.

SIMULATION MODE

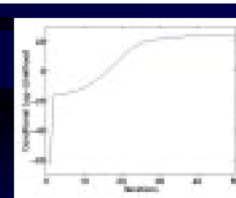
Fully synthesize both components of the time series (user A and user B). Some instabilities / bugs (no grounding) -> future work.

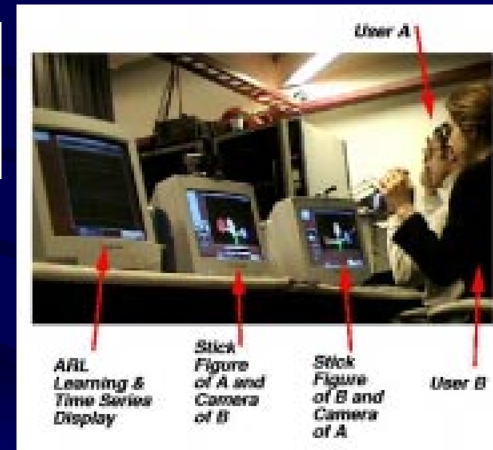


Training & Results

Training Data: 5 minutes, 15 Hz, approx. 5 repetitions of gestures
Model: T=120, Dims=22, M=25 (to maintain real-time)
Convergence: 2 hours

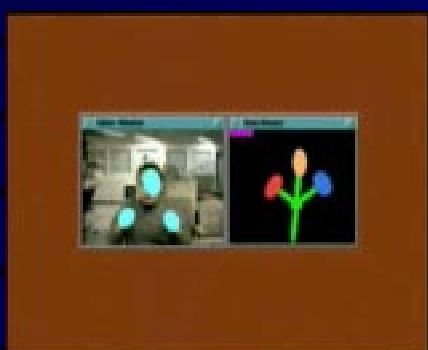
Interaction	User	Corresponding Action
1	A	Scare B by moving towards camera
	B	Fearfully crouch down & bring hands in
2	A	Wave hello
	B	Wave back accordingly
3	A	Circle stomach & tap head
	B	Clap enthusiastically
4	A	Idle or Small Gestures
	B	Idle or Small Gestures





Nearest Neighbor: 1.57% RMS
Constant Velocity: 0.85% RMS
ARL: 0.64% RMS

Results...

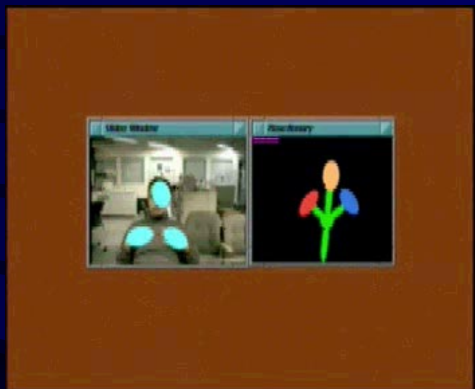


SCARE INTERACTION

Results...

WAVE INTERACTION

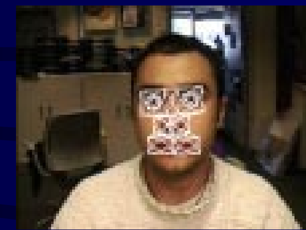
Results...

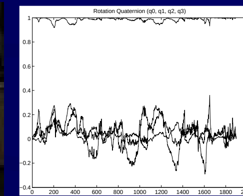


CLAP INTERACTION

Alternate Perceptual Modalities

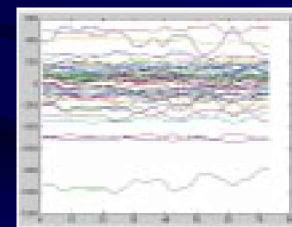
FACE MODELING

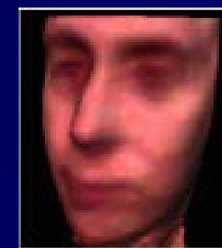




3D Translation
3D Pose
Focal Length
7 DOFs

3D Eigen Model
Deformations
Texture
40 DOFs





Real-Time
(…+ speech +…)

Conclusions

1 - Unsupervised Discovery of Simple Interactive Behaviour by Perceptual Observations and Statistical Time Series Prediction of Future Given Past or Reaction Given Action

2 - Imitation Based Learning of Behaviour

3 - Real-Time Behaviour Acquisition and Interactive Synthesis

4 - Small Amounts of Training Data and Non-Intrusive Training

5 - Non-Linear Predictive Model for Feedback Perception

6 - Monotonically Convergent Maximum Conditional Likelihood i.e. Discriminant Probabilistic Learning (CEM)

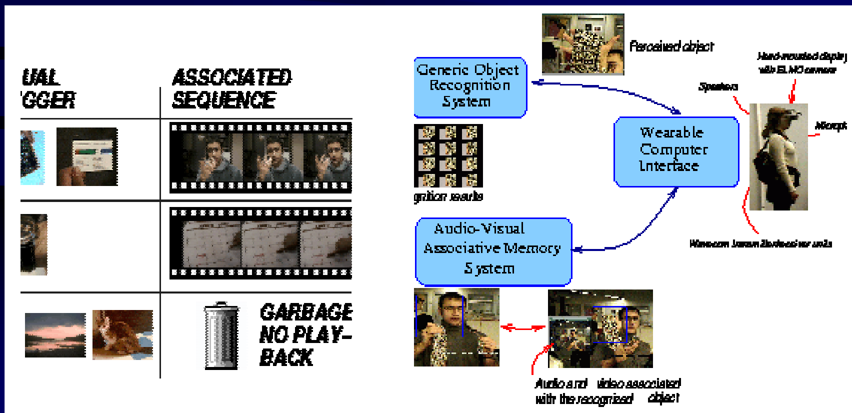
7 - No A Priori Segmentation or Classification of Gestures

Wearable Platform: Dynamic Personal Enhanced Reality System

Tony Jebara
Bernt Schiele
Nuria Oliver
Alex Pentland

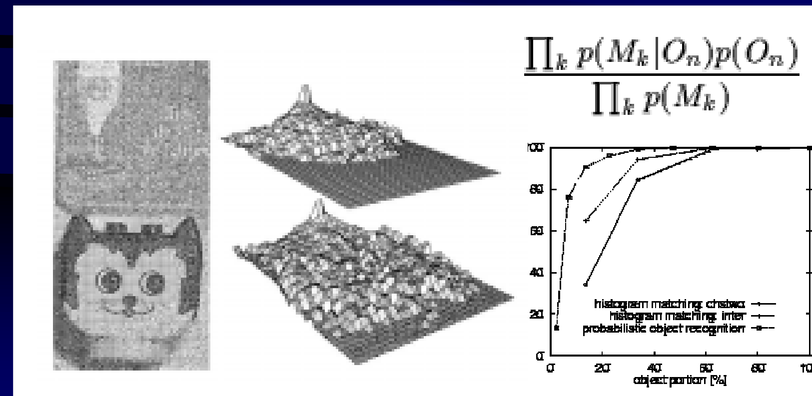
DyPERS Architecture

* 3 Button interface: Record, Associated and Discard
* User records live A/V Clips with wearable
* Associates them with a visual trigger object(s)
* Audio-Video is replayed when computer vision sees trigger object

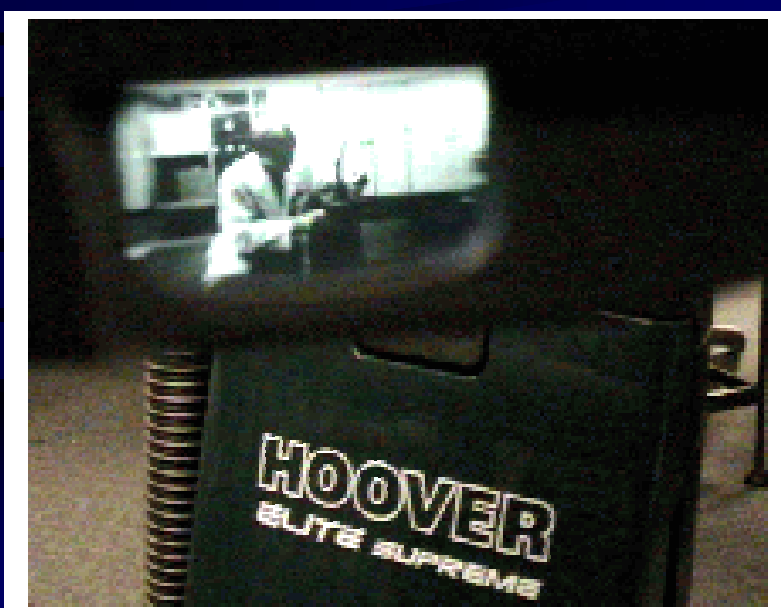


DyPERS Visual Recognition

* Multidimensional filter-response histograms
  from differences of Gaussian linear convolutions
  (magnitude of 1st deriv. and Laplace operator)
* Compute probability of object from k iid measurements
* Kalman filter on probabilities of objects to smooth classifier



Video



Wearable Platform: Interactive Behavior Acquisition

Tony Jebara
Alex Pentland

Wearable Long-term Behavior Acquisition

Hardware
-Sony Picturebook Laptop
-2 Cameras
-2 Microphones

Action-Reaction Learning
-Audio-Visual processing
-Time Series Prediction
-Discriminative Learning
-DyPERS associative memory

Tracking Conversational Context (Audio)

Tony Jebara
Yuri Ivanov
Ali Rahimi
Alex Pentland

System Architecture

Tracking Conversational Context for Machine Mediation
Speech rec. on multiple speakers with real-time topic-spotting

Bag-of-words (multinomials)

$$P(word_i|c) = \frac{N_c(word_i)}{\sum_{j=1}^{w} N_c(word_j)}$$

Short Term Memory w/ Decay

$$x_i^t = \alpha x_i^{t-1} + \delta(k,i)$$

Probability of Topic

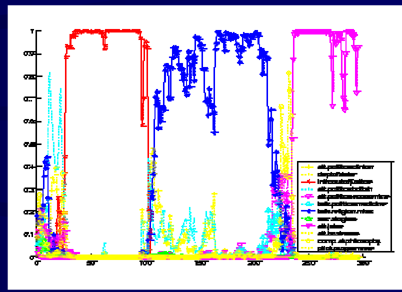
$$P(\mathbf{x}|c) = \prod_i P(word_i|c)^{x_i}$$

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{\sum_{k=1}^{c} P(\mathbf{x}|k)P(k)}$$

Coarse descriptor of mood, topic, etc. Used to select prompt to stimulate conversation

Conversation Trace

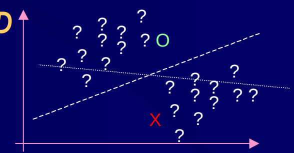
Text data from 12 Newsgroups



Real-time probabilities of topics (politics, health, religion, etc.)
Could also detect emotions & situations...

Wearable Platform

Clutering & MED Transductive approach for few labeled.



References

1) Tony Jebara and Alex Pentland. On Reversing Jensen's Inequality. In *Neural Information Processing Systems 13 (NIPS*'00)* , Dec. 2000.

2) Tony Jebara, Yuri Ivanov, Ali Rahimi and Alex Pentland. Tracking Conversational Context for Machine Mediation of Human Discourse. In *AAAI Fall 2000 Symposium - Socially Intelligent Agents - The Human in the Loop*. Nov. 2000.

3) Tony Jebara and Tommi Jaakkola. Feature Selection and Dualities in Maximum Entropy Discrimination. In *16th Conf. Uncertainty in Artificial Intelligence (UAI 2000)*.

4) Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum Entropy Discrimination. In *Neural Information Processing Systems 12 (NIPS*'99)*.

5) Tony Jebara and Alex Pentland. Action Reaction Learning: Automatic Visual Analysis and Synthesis of Interactive Behaviour. In *1st Intl. Conf. on Computer Vision Systems (ICVS'99)*.

6) Bernt Schiele, Nuria Oliver, Tony Jebara and Alex Pentland. An Interactive Computer Vision System, DyPERS: Dynamic Personal Enhanced Reality System. In *1st Intl. Conf. on Computer Vision Systems (ICVS'99)*.

7) Tony Jebara and Alex Pentland. Maximum Conditional Likelihood via Bound Maximization and the CEM Algorithm. In *Neural Information Processing Systems 11 (NIPS*'98)* .

8) Tony Jebara. Action-Reaction Learning: Analysis and Synthesis of Human Behaviour. Master's Thesis, MIT, May 1998.