

Advanced Machine Learning & Perception

Instructor: Tony Jebara

Maximum Entropy Discrimination

- Combining Generative and Discriminative
- The Exponential Family
- Maximum Entropy
- Minimum Relative Entropy
- Support Vector Machines Revisited
- Maximum Entropy Discrimination
- Discriminative Generative Models

The Exponential Family

- When are integrals or ML, MAP “nice” mathematically?
- Can we deal with other types of data than Gaussian?
- Yes! The general family of such distributions (includes many classics) is the **Exponential Family** which has the following “natural form”:

$$p(x | \theta) = h(x) \exp(\theta^T T(x) - A(\theta))$$

model and data interact only via dot product

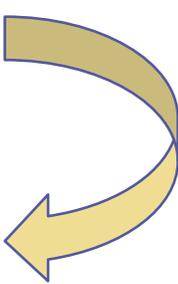
the **cumulant generating function** $A(\theta)$ is convex

- Examples of E-Family:
 - Bernoulli: binary
 - Multinomial: discrete
 - Gaussian: real vectors
 - Poisson: positive integers
 - Dirichlet: positive reals

Natural Parameter Form

• Gaussian Natural Form: $p(x | \theta) = h(x) \exp(\theta^T T(x) - A(\theta))$

$$\begin{aligned}
 p(x | \mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \\
 &= \frac{1}{\sqrt{2\pi}} \exp\left(-\log \sigma - \frac{1}{2\sigma^2}(x - \mu)^2\right) \\
 &= \frac{1}{\sqrt{2\pi}} \exp\left(-\log \sigma - \frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right)
 \end{aligned}$$



$$\theta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi}} \exp\left(T_1\theta_1 + T_2\theta_2 - \log \sigma - \frac{1}{2\sigma^2}\mu^2\right)
 \end{aligned}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$A(\theta) = \frac{\mu^2}{2\sigma^2} + \log \sigma = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2} \log(-2\theta_2)$$

• If multidimensional: $\theta = \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix}$ $T(x) = \begin{bmatrix} x \\ xx^T \end{bmatrix}$

Properties of E-Family

•It's normalized: $\int_x p(x | \theta) dx = \int_x h(x) \exp(\theta^T T(x) - A(\theta)) dx$

$$1 = \int_x h(x) \exp(\theta^T T(x)) dx \exp(-A(\theta))$$

A(θ) is Laplace Transform of h(x)

$$A(\theta) = \log \int_x h(x) \exp(\theta^T T(x)) dx$$

1st derivative

$$\begin{aligned} \frac{\partial A(\theta)}{\partial \theta} &= \frac{\int_x h(x) \exp(\theta^T T(x)) T(x) dx}{\int_x h(x) \exp(\theta^T T(x)) dx} \\ &= \frac{\int_x h(x) \exp(\theta^T T(x) - A(\theta)) T(x) dx}{\int_x h(x) \exp(\theta^T T(x) - A(\theta)) dx} \\ &= E_{p(x|\theta)} \{ T(x) \} \end{aligned}$$

2nd derivative

$$\begin{aligned} \frac{\partial^2 A(\theta)}{\partial \theta} &= \frac{\int_x h(x) \exp(\theta^T T(x)) T^2(x) dx}{\int_x h(x) \exp(\theta^T T(x)) dx} \\ &\quad - \left(\frac{\int_x h(x) \exp(\theta^T T(x)) T(x) dx}{\int_x h(x) \exp(\theta^T T(x)) dx} \right)^2 \\ &= E \{ T^2(x) \} - E \{ T(x) \}^2 \\ &= Cov_{p(x|\theta)} \{ T(x) \} \end{aligned}$$

Maximum Likelihood E-Family

- Maximum Likelihood (IID) with E-Family:

all ML problems are easy!

$$l(\theta) = \sum_i \log p(x_i | \theta)$$

$$l(\theta) = \sum_i \log h(x_i) \exp(\theta^T T(x_i) - A(\theta))$$

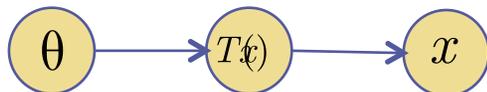
$$l(\theta) = \sum_i \log h(x_i) + \theta^T T(x_i) - A(\theta)$$

$$l'(\theta) = \sum_i T(x_i) - NA'(\theta)$$

$$0 = \sum_i T(x_i) - NA'(\theta)$$

$$A'(\theta) = \frac{1}{N} \sum_i T(x_i)$$

- Note $A(\theta)$ is convex (Cov is positive semi-definite)
- Convex means gradient is unique so *global* ML solution
- **Sufficient Statistics:** (Pitman, Koopman '36) summarize data via simple averages over finite functions of it $T(x)$



$$A'(\theta) = E_{p(x|\theta)} \{T(x)\} = \frac{1}{N} \sum_i T(x_i)$$

Conjugate Priors of E-Family

- Each e-family has dual or conjugate distrib over its params

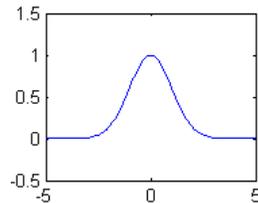
$$p(x | \theta) = h(x) \exp\left(R(\theta)^T T(x) - A(\theta)\right)$$

$$p(\theta | X) = j(n, X) \exp\left(R(\theta)^T T(x) - nA(\theta)\right)$$

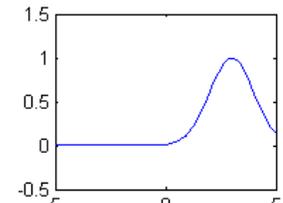
Like n virtual data points

- Conjugates:

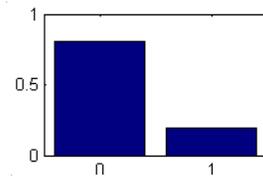
Gaussian Mean



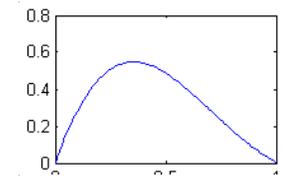
Gaussian Mean



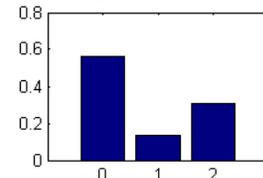
Bernoulli



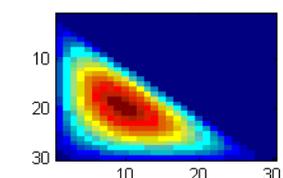
Beta



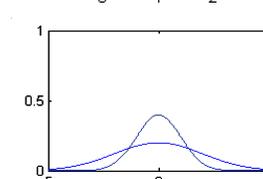
Multinomial



Dirichlet



Gaussian Cov



Inverse Wishart

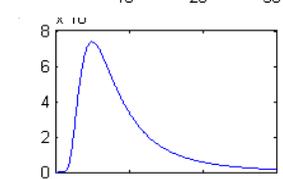


Table of E-Families

- Examples of the $H()$ and $A()$ functions for E-Families...

$$p(x | \theta) = h(x) \exp(\theta^T T(x) - A(\theta))$$

$$= \exp\left(H(x) + \theta^T T(x) - A(\theta)\right)$$

	$H(x)$	$A(\theta)$	<i>domain of θ</i>
Gaussian	$-\frac{D}{2} \log(2\pi)$	$\frac{1}{2} \log(-2\theta_2^{-1}) - \frac{1}{4} \theta_1^T \theta_2^{-1} \theta_1$	$\theta_1 \in \mathbb{R}^D$ $\theta_2 \in \mathbb{R}^{D,D} < 0$
Multinomial	$\log(\Gamma(\eta + 1)/\nu)$ $\eta = \sum_{d=1}^{D+1} X_d$ $\nu = \prod_d \Gamma(X_d + 1)$	$\eta \log(1 + \sum_{d=1}^D \exp(\Theta_d))$	$\Theta \in \mathbb{R}^D$
Exponential	0	$-\log(-\Theta)$	$\Theta \in \mathbb{R}_-$
Gamma	$-\exp(X) - X$	$\log \Gamma(\Theta)$	$\Theta \in \mathbb{R}_+$
Poisson	$\log(X!)$	$\exp(\Theta)$	$\Theta \in \mathbb{R}$

- But, we can also *derive* e-family form via Maximum Entropy

Maximum Entropy

- Maximum Entropy Theory (Jaynes, Shannon, 1950-1960's)

Find a distribution $p(x)$ which satisfies certain constraints while maximizing the entropy of $p(x)$

- Based on principle of indifference or insufficient reason

- Entropy of $p(x)$ is: $H(p) = -\sum_x p(x) \log p(x)$

- Maximum entropy wants to spread probability on all x

- Uniform distribution has highest H , $p = \arg \max H(p)$

- But, let's say we add a few "expectation" constraints:

$$E_{p(x)} \{f_i(x)\} = \sum_x p(x) f_i(x) = \alpha_i$$

- Under $p(x)$, force expected $f(x)$ value to be a constant α

- Examples: $f(x) = x$ constrains the mean
 $f(x) = x^p$ constrains the higher p moments
 $f(x) = 1$ constant function (what is α ?)

MaxEnt & Min Relative Entropy

- User gives $f(x)$ feature or moment functions and α values

$$E_{p(x)} \{f_1(x)\} = \sum_x p(x) f_1(x) = \alpha_1$$

$$E_{p(x)} \{f_N(x)\} = \sum_x p(x) f_N(x) = \alpha_N$$

$f(x)$'s can be
any function

we also ALWAYS have the $f_0(x)=1$ and $\alpha_0=1$

- These constraints select out a piece \mathfrak{P} of probability space
- Want maximum entropy $p(x)$ in it $p = \arg \max_{p \in \mathfrak{P}} H(p)$
- Instead of maximizing entropy, we can equivalently minimize relative entropy to the uniform distribution:
- Note Kullback-Leibler Divergence (Relative Entropy)

$$KL(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$KL(p \parallel u) = \sum_x p(x) \log p(x) - \sum_x p(x) C$$

**If $q(x)$ is uniform
this is just $-H(p)$**

Minimum Relative Entropy

- So, more generally, we can solve maximum entropy as a minimum relative entropy problem where q is uniform
- The general problem now is: $p = \arg \min_{p \in \mathfrak{P}} KL(p \parallel q)$
- Represent the constraints on p with Lagrange multipliers to obtain the following primal Lagrangian to minimize:

$$\mathcal{L} = KL(p \parallel q) - \sum_i \lambda_i \left(\sum_x p(x) f_i(x) - \alpha_i \right) - \gamma \left(\sum_x p(x) - 1 \right)$$

- Since KL is convex and constraints are linear, can solve by taking derivatives with respect to a $p(x)$ and setting to 0:

$$1 + \log p(x) - \log q(x) - \sum_i \lambda_i f_i(x) - \gamma = 0$$

$$p(x) = q(x) \exp\left(\gamma - 1 + \sum_i \lambda_i f_i(x)\right)$$

$$p(x) = \frac{1}{Z} q(x) \exp\left(\sum_i \lambda_i f_i(x)\right)$$

Isolate $p(x)$

Gamma is given by normalization Z

Duality & the Partition Function

- The normalizer or partition function Z depends on lambdas

$$Z(\lambda) = \sum_x q(x) \exp\left(\sum_i \lambda_i f_i(x)\right)$$
- Maximum entropy solution is in exponential family form!

$$p(x) = q(x) \exp\left(\sum_i \lambda_i f_i(x) - \log Z(\lambda)\right)$$
- Here, λ Lagranges behave like θ parameters
- We still haven't specified the values for λ ...
- Reinsert above solution for $p(x)$ into primal Lagrangian:

$$\begin{aligned} \mathcal{L} &= KL(p \| q) - \sum_i \lambda_i \left(\sum_x p(x) f_i(x) - \alpha_i\right) - \gamma \left(\sum_x p(x) - 1\right) \\ &= \sum_x p(x) \log \frac{\exp(\sum_i \lambda_i f_i(x))}{Z(\lambda)} - \sum_i \lambda_i \left(\sum_x p(x) f_i(x) - \alpha_i\right) \\ &= -\log Z(\lambda) + \sum_i \lambda_i \alpha_i \end{aligned}$$
- Get λ scalars by *maximizing* the concave dual Lagrangian:

$$\mathcal{D} = -\log Z(\lambda) + \sum_i \lambda_i \alpha_i$$

Dual Solution

- Final scalar λ 's found by maximizing $-\log Z(\lambda) + \sum_i \lambda_i \alpha_i$ and inserted into: $p(x) = \frac{1}{Z(\lambda)} q(x) \exp\left(\sum_i \lambda_i f_i(x)\right)$
- Final solution satisfies constraints (and maxes likelihood):

$$\sum_x p(x) f_i(x) = \alpha_i \approx \frac{1}{T} \sum_{t=1}^T f_i(x_t) \text{ where } x_t \sim p(x)$$

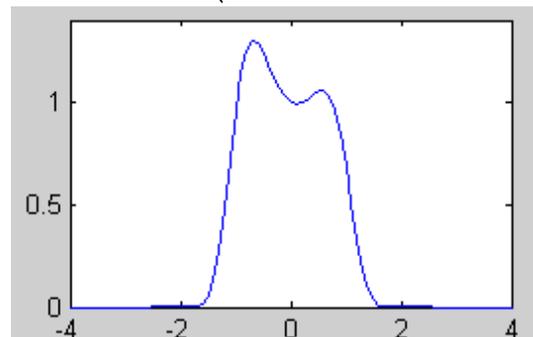
- Thus, the maximum entropy practitioner designs the constraints on the problem, not the parametric form

$$E_{p(x)} \{x\} = -0.0777 \quad \Rightarrow \quad p(x) \propto \exp(-.16x - .21x^2 + x^3 - x^4)$$

$$E_{p(x)} \{x^2\} = 0.4767$$

$$E_{p(x)} \{x^3\} = -0.0719$$

$$E_{p(x)} \{x^4\} = -0.4422$$



Maximum Entropy Inequalities

- Unlike EM (which finds the previous distribution with a mix of 2 Gaussians), Maximum Entropy is always unique globally guaranteed (but may have v. many $f(x)$ functions)
- We can also consider *inequality* constraints:

$$\begin{aligned} \sum_x p(x) f_1(x) &= \alpha_1 \\ \sum_x p(x) f_2(x) &\geq \alpha_2 \\ \sum_x p(x) f_3(x) &\leq \alpha_3 \end{aligned}$$

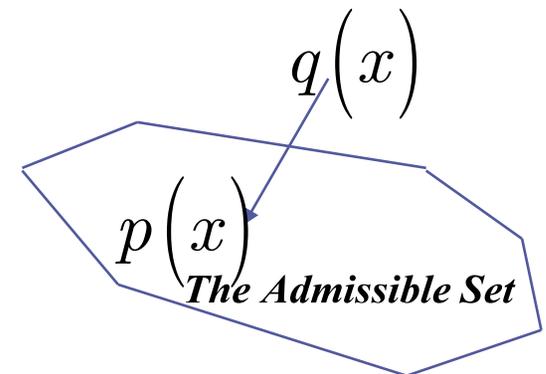
**Corresponding
Lagrange
multipliers
become +ve
or -ve only**

$$\lambda_1 \in \mathbb{R}$$

$$\lambda_2 \in \mathbb{R}^+$$

$$\lambda_3 \in \mathbb{R}^-$$

- In general, these linear constraints carve up a piece \mathfrak{P} of probability space with lines or half-planes
- Min relative entropy finds closest distribution to hull via an information geometric projection

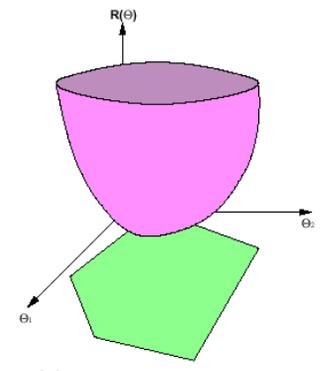
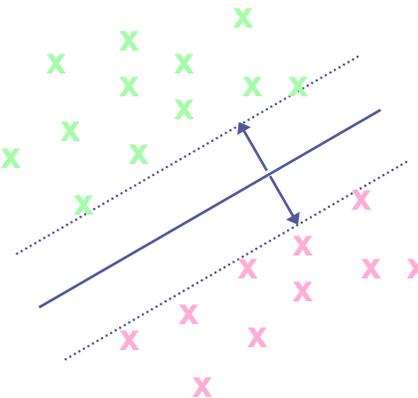


Back to SVMs & Regularization

- Given: training examples: $\{X_1, \dots, X_T\}$
 binary (+/- 1) labels: $\{y_1, \dots, y_T\}$
 discriminant function: $L(X; \Theta) = \theta^T X + b$
 where: $\Theta = \{\theta, b\}$
- Maximize margin or min penalty/regularizer function: $\frac{1}{2} \|\theta\|^2$
 subject to classification constraints:

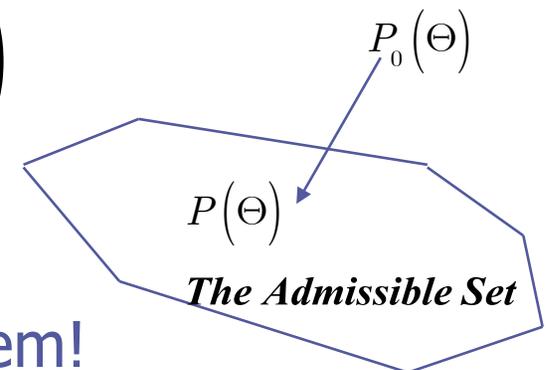
$$y_t L(X_t; \Theta) - 1 \geq 0, \forall t$$

- Solve using QP, Get best Θ^*
- Gives max margin hyperplane
- But, not probabilistic
- Not flexible & generative
- No priors, structure, etc.



SVMs via Maximum Entropy

- Support vector machine finds only the single best Θ^*
- But, other solutions close to Θ^* may be almost as good
- Instead consider solving for a distribution $P(\Theta)$
- Put high probability weight on good solutions
- Can get back same answer as Θ^* if we set $P(\Theta) = \delta(\Theta - \Theta^*)$
- How to solve for $P(\Theta)$? Satisfy classification constraints in the *expected sense*: $\int P(\Theta) [y_t L(X_t; \Theta) - 1] d\Theta \geq 0, \forall t$
- What is the regularization penalty function? use KL to user given prior: $KL(P \parallel P_0)$
- Also *use* the classifier in expectation: $\hat{y} = \text{sign} \int P(\Theta) L(X; \Theta) d\Theta$
- Above is now a maximum entropy problem!



Maximum Entropy Discrimination

- Maximum entropy solution as usual, but now over Θ not X :

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

- Compute the partition function (just continuous integral)

$$Z(\lambda) = \int_{\Theta} P_0(\Theta) \exp\left(\sum_t \lambda_t \left[y_t L(X_t; \Theta) - 1\right]\right) d\Theta$$

- Maximize the dual negated log of it as before (alphas?)

$$J(\lambda) = -\log Z(\lambda) \quad \text{where} \quad \lambda = \{\lambda_1, \dots, \lambda_T\}$$

- Have 1 constraint per data point instead of per moment fn
our feature functions $f_i(\Theta)$ now become:

$$f_t(\Theta) = y_t L(X_t; \Theta) - 1$$

- We don't want a certain mean, variance, etc. for $P(\Theta)$
- Our constraints are that it gives a good classifier
- Let us see the solution for this maximum entropy problem

Maximum Entropy Discrimination

- We still need to specify a prior $P_0(\Theta)$, use a zero mean identity-covariance Gaussian on linear vector and another zero-mean Gaussian on the bias:

$$P_0(\Theta) = P_0(\theta, b) = N(\theta | 0, I) N(b | 0, \sigma^2)$$

$$L(X; \Theta) = \theta^T X + b \quad \text{recall}$$

- Now, compute the partition function:

$$\begin{aligned} Z(\lambda) &= \int_{\Theta} P_0(\Theta) \exp\left(\sum_t \lambda_t \left[y_t L(X_t; \Theta) - 1\right]\right) d\Theta \\ &= \int_{\theta} N(\theta | 0, I) \exp\left(\sum_t \lambda_t y_t \theta^T X_t\right) d\theta \\ &\quad \times \int_b N(b | 0, \sigma^2) \exp\left(\sum_t \lambda_t y_t b\right) db \times \exp\left(-\sum_t \lambda_t\right) \\ &= \int_{\theta} \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2} \theta^T \theta + \theta^T \sum_t \lambda_t y_t X_t\right) d\theta \times \dots \\ &= \int_{\theta} \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2} \theta^T \theta + \theta^T Q - \frac{1}{2} Q^T Q + \frac{1}{2} Q^T Q\right) d\theta \times \dots \end{aligned}$$

Q

MED: Gives Back SVMs

- Continuing our partition function:

$$\begin{aligned} Z(\lambda) &= \int_{\theta} \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}\theta^T\theta + \theta^T Q - \frac{1}{2}Q^T Q + \frac{1}{2}Q^T Q\right) d\theta \times \dots \\ &= \exp\left(\frac{1}{2}Q^T Q\right) \int_b N(b \mid 0, \sigma^2) \exp\left(\sum_t \lambda_t y_t b\right) db \times \exp\left(-\sum_t \lambda_t\right) \\ &= \exp\left(\frac{1}{2}Q^T Q\right) \exp\left(\frac{1}{2}\sigma^2 \left(\sum_t \lambda_t y_t\right)^2\right) \exp\left(-\sum_t \lambda_t\right) \end{aligned}$$

- Now, compute the objective as $J = -\log(Z)$ to maximize

$$\begin{aligned} J(\lambda) &= -\log Z(\lambda) = -\frac{1}{2}Q^T Q - \frac{1}{2}\sigma^2 \left(\sum_t \lambda_t y_t\right)^2 + \sum_t \lambda_t \\ &= \sum_t \lambda_t - \frac{1}{2}\sigma^2 \left(\sum_t \lambda_t y_t\right)^2 - \frac{1}{2} \sum_{t,t'} \lambda_t \lambda_{t'} y_t y_{t'} \left(X_t^T X_{t'}\right) \end{aligned}$$

- Assume non-informative prior (all b values are equal), then σ goes to infinity. Thus, for J to be high, need $\sum_t \lambda_t y_t = 0$

$$J(\lambda) = \sum_t \lambda_t - \frac{1}{2} \sum_{t,t'} \lambda_t \lambda_{t'} y_t y_{t'} \left(X_t^T X_{t'}\right) \text{ s.t. } \sum_t \lambda_t y_t = 0, \lambda_t \geq 0 \forall t$$

SVM AGAIN!

WHY +VE?

MED: Nonlinear Classifiers?

• We can now consider nonlinear discriminant functions *without* using the kernel trick...

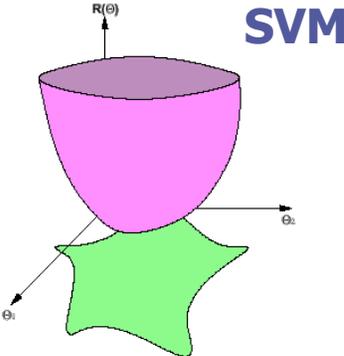
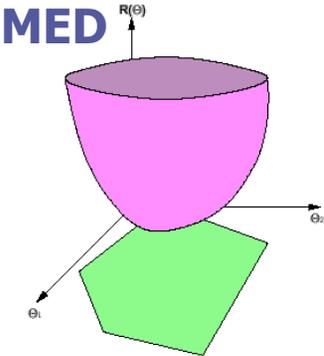
• For example a generative-style classifier discriminant function from the ratio of two probability models:

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

- Full Gaussians
- Tree Structures
- Bernoulli
- Multinomials
- Exponentials
- HMMs
- Mixtures

• Entropy approach stays convex

$$\int P(\Theta) y_t L(X_t; \Theta) \geq 0$$



$$y_t L(X_t; \Theta) \geq 0$$

SVM breaks

MED: E-Family Ratio Classifiers

- When can we actually solve for $L(X; \Theta) = \log \frac{P(X|\theta_+)}{P(X|\theta_-)} + b$
- If ratio of exponential families and use conjugates for priors

$$p(X | \theta) = \exp(A(X) + X^T \theta - K(\theta))$$

$$p(\theta | \chi) = \exp(\tilde{A}(\theta) + \theta^T \chi - \tilde{K}(\chi))$$

- Proof: Partition function analytically computable

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

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

Again Use
Non-Informative
Prior on b

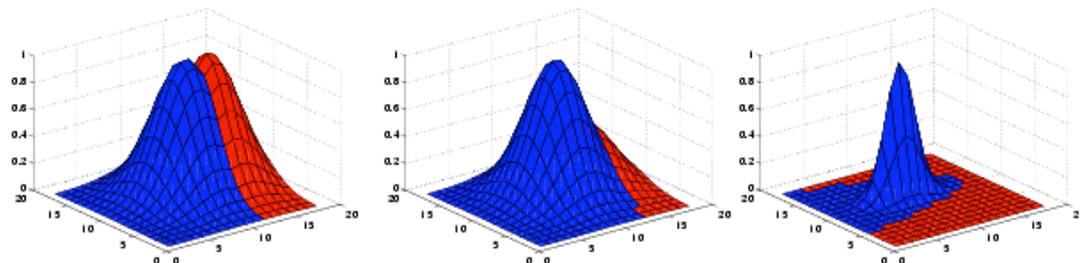
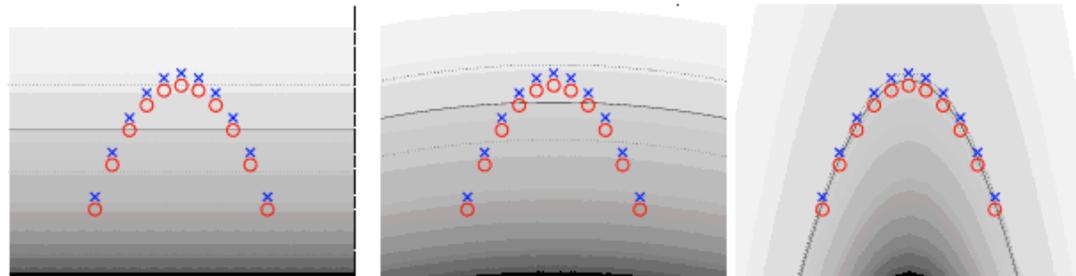
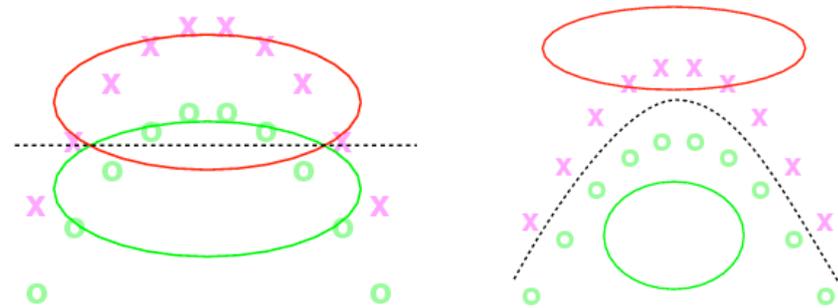
$$Z_{\theta_{\pm}} = \int \exp(\tilde{A}(\theta_{\pm}) + \theta_{\pm}^T \chi - \tilde{K}(\chi)) \exp\left(\sum_t \lambda_t y_t (A(X_t) + X_t^T \theta_{\pm} - K(\theta_{\pm}))\right) d\theta_{\pm}$$

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

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

MED: Ratio of Gaussians

$$L(X; \Theta) = \log \frac{N(X|\mu_+, \Sigma_+)}{N(X|\mu_-, \Sigma_-)} + b$$



(a) ML & MED Initialization

(b) MED Intermediate

(c) MED Converged

MED: Ratio of Gaussians

UCI Breast
Cancer
Data Set:

Method	Training Errors	Testing Errors
Nearest Neighbor		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

UCI Crabs
Data Set:

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

MED: Ratio of Gaussians

