Advanced Machine Learning & Perception

Instructor: Tony Jebara

Maximum Entropy

- 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":

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

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

$$p\left(x \mid \mu,\sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\log\sigma - \frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\log\sigma - \frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right)$$

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

$$\theta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \qquad T\left(x\right) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$h\left(x\right) = \frac{1}{\sqrt{2\pi}} \qquad A\left(\theta\right) = \mu^2/2\sigma^2 + \log\sigma = -\theta_1^2/4\theta_2 - \frac{1}{2}\log\left(-2\theta_2\right)$$

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

Properties of E-Family

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

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

A(θ) is Laplace Transform of h(x)

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

1st derivative

$$\frac{\partial A(\theta)}{\partial \theta} = \frac{\int_x h(x) \exp\left(\theta^T T(x)\right) T(x) dx}{\int_x h(x) \exp\left(\theta^T T(x)\right) dx}$$

$$= \frac{\int_x h(x) \exp\left(\theta^T T(x) - A(\theta)\right) T(x) dx}{\int_x h(x) \exp\left(\theta^T T(x) - A(\theta)\right) dx}$$

$$= E_{p(x|\theta)}\left\{T(x)\right\}$$

2nd derivative

$$\frac{\partial^2 A(\theta)}{\partial \theta} = \frac{\int_x h(x) \exp\left(\theta^T T(x)\right) T^2(x) dx}{\int_x h(x) \exp\left(\theta^T T(x)\right) dx}$$

$$- \left(\frac{\int_x h(x) \exp\left(\theta^T T(x)\right) T(x) dx}{\int_x h(x) \exp\left(\theta^T T(x)\right) dx} \right)^2$$

$$= E\left\{T^2(x)\right\} - E\left\{T(x)\right\}^2$$

$$= Cov_{p(x|\theta)}\left\{T(x)\right\}$$

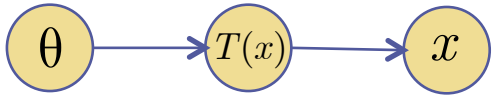
Maximum Likelihood E-Family

- Maximum Likelihood (IID) with E-Family:

$$l(\theta) = \sum_i \log p(x_i \mid \theta)$$
$$l(\theta) = \sum_i \log h(x_i) \exp\left(\theta^T T(x_i) - A(\theta)\right)$$
$$l(\theta) = \sum_i \log h(x_i) + \theta^T T(x_i) - A(\theta)$$
$$l'(\theta) = \sum_i T(x_i) - NA'(\theta)$$
$$0 = \sum_i T(x_i) - NA'(\theta)$$
$$A'(\theta) = \frac{1}{N} \sum_i T(x_i)$$

all ML problems are easy!

- Note A(θ) 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)



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

Conjugate Priors of E-Family

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

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

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

Like n virtual data points

- Conjugates:

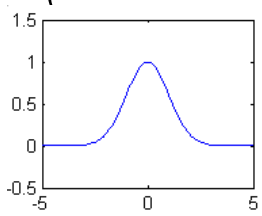
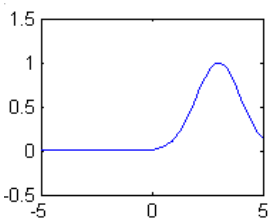
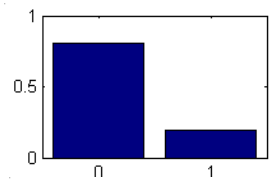
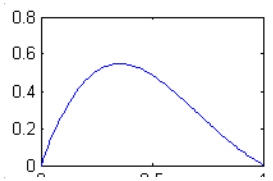
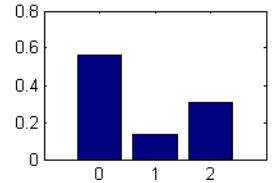
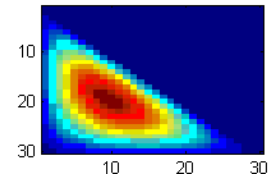
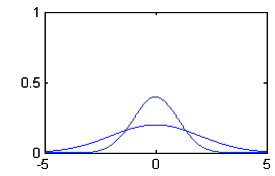
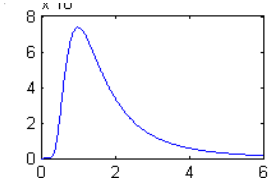
Gaussian Mean 	Gaussian Mean 
Bernoulli 	Beta 
Multinomial 	Dirichlet 
Gaussian Cov 	Inverse Wishart 

Table of E-Families

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

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

	$H\left(x\right)$	$A\left(\theta\right)$	$domain\ of\ \theta$
Gaussian	$-\frac{D}{2}\log(2\pi)$	$\frac{1}{2}\log(\lvert -2\theta_2^{-1}\rvert) - \frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1$	$\theta_1 \in \mathbb{R}^D$ $\theta_2 \in \mathbb{R}^{D,D} < 0$
Multinomial	$\log(\Gamma(\eta+1)/\nu)$ $\eta = \sum_{d=1}^{D+1} X_d$ $\nu = \Pi_d \Gamma(X_d+1)$	$\eta\ \log(1+\sum_{d=1}^{D}\exp(\Theta_d))$	$\Theta \in \mathbb{R}^D$
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: $H\left(p\right) = -\sum_x p\left(x\right)\log p\left(x\right)$
- Maximum entropy wants to spread probability on all x
- Uniform has highest H=log(N), $u = \arg\max_p H\left(p\right)$
- But, let's say we add a few "expectation" constraints:

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

- 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

$$E_{p(x)}\left\{f_1(x)\right\} = \sum_x p(x) f_1(x) = \alpha_1$$
$$\cdots$$
$$E_{p(x)}\left\{f_N(x)\right\} = \sum_x p(x) f_N(x) = \alpha_N$$

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)

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$
$$KL(p\|u) = \sum_x p(x) \log p(x) - \sum_x p(x) C$$

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: $p = \arg\min_{p \in \mathfrak{P}} KL\left(p \middle\| q\right)$
- Represent the constraints on p with Lagrange multipliers to obtain the following primal Lagrangian to minimize:

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

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

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

Isolate p(x)

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

Gamma is given by normalization Z

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

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:

$$\begin{aligned}
\mathfrak{L} &= \max_{\lambda,\gamma} \min_p KL\left(p \| q\right) - \sum_i \lambda_i \left(\sum_x p(x) f_i(x) - \alpha_i\right) - \gamma\left(\sum_x p(x) - 1\right) \\
&= \max_\lambda \sum_x p(x) \log \frac{\exp\left(\sum_i \lambda_i f_i(x)\right)}{Z(\lambda)} - \sum_i \lambda_i \left(\sum_x p(x) f_i(x) - \alpha_i\right) \\
&= \max_\lambda - \log Z(\lambda) + \sum_i \lambda_i \alpha_i
\end{aligned}$$

- Get λ scalars by *maximizing* the concave dual Lagrangian:
$$\mathfrak{D} = -\log Z(\lambda) + \sum_i \lambda_i \alpha_i$$

Dual Solution

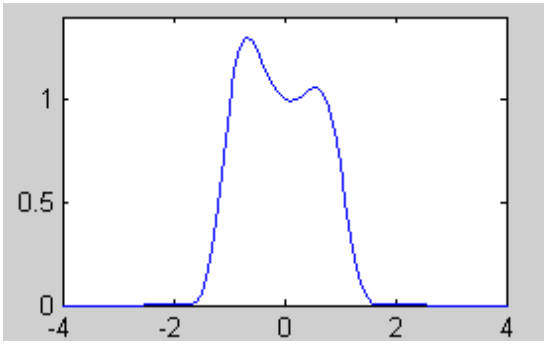
- Final scalar λ's found by maximizing $-\log Z\left(\lambda\right)+\sum_i \lambda_i \alpha_i$ and inserted into: $p\left(x\right)=\frac{1}{Z\left(\lambda\right)} q\left(x\right)\exp\left(\sum_i \lambda_i f_i\left(x\right)\right)$
- Final solution satisfies constraints (and maxes likelihood):

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

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

$$E_{p(x)}\left\{x\right\}=-0.0777$$
$$E_{p(x)}\left\{x^2\right\}=0.4767$$
$$E_{p(x)}\left\{x^3\right\}=-0.0719$$
$$E_{p(x)}\left\{x^4\right\}=-0.4422$$

$$\Rightarrow \quad p\left(x\right)\propto \exp\left(-.16x-.21x^2+x^3-x^4\right)$$



Maximum Entropy Inequalities

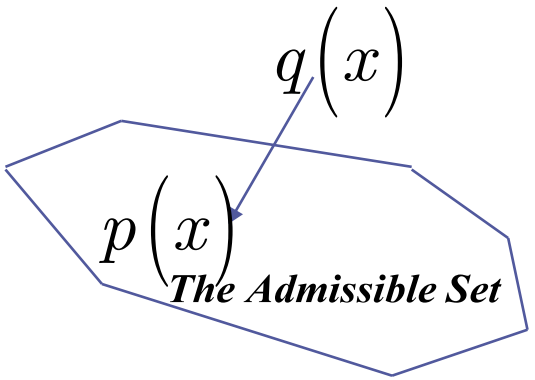
- Unlike EM (which finds the previous distribution with a mix of 2 Gaussians), Maximum Entropy is always unique globally guaranteed (but may have v. many f(x) functions)
- We can also consider *inequality* constraints:

$$\sum_x p(x) f_1(x) = \alpha_1$$
$$\sum_x p(x) f_2(x) \geq \alpha_2$$
$$\sum_x p(x) f_3(x) \leq \alpha_3$$

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

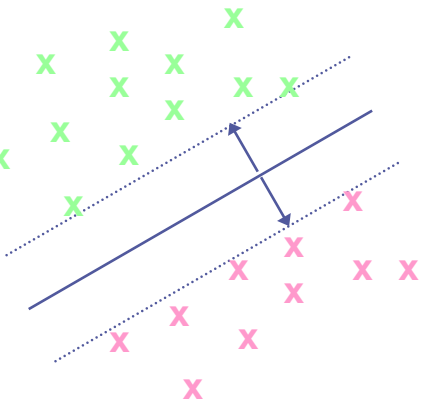


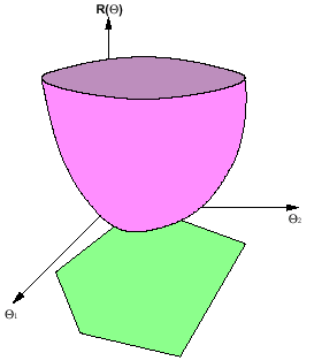
Recall Support Vector Machines

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

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

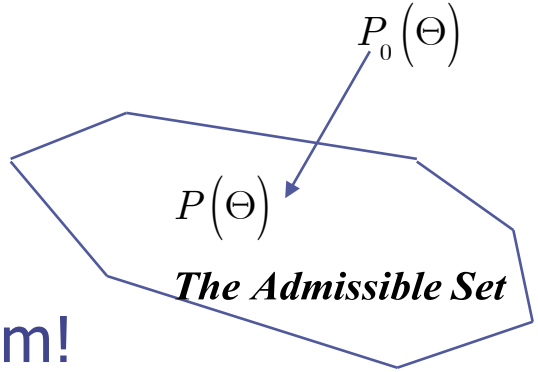
- Solve using QP, Get best Θ^*
- Gives max margin hyperplane
- But, not probabilistic
- Not flexible & generative
- No priors, structure, etc.





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*: $\int P(\Theta)\left[y_t f(X_t;\Theta) - 1\right] d\Theta \geq 0, \ \forall t$
- What is the regularization penalty function? use KL to user given prior: $KL(P \| P_0)$
- Also *use* the classifier in expectation: $\hat{y} = sign \int P(\Theta) f(X;\Theta) d\Theta$
- Above is now a maximum entropy problem!



Maximum Entropy Discrimination

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

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

- Compute the partition function (just continuous integral)

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

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

$$J(\lambda) = -\log Z(\lambda) \qquad \text{where} \qquad \lambda = \left\{\lambda_1, \ldots, \lambda_T\right\}$$

- Have 1 constraint per data point instead of per moment fn

  our feature functions $f_i(\Theta)$ now become:

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

- 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:

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

recall $f(X;\Theta) = \theta^T X + b$

- Now, compute the partition function:

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

Q

MED: Gives Back SVMs

- Continuing our partition function:

$$Z(\lambda) = \int_{\theta} \frac{1}{(2\pi)^{D/2}} \exp\left(-\tfrac{1}{2}\theta^T\theta + \theta^T Q - \tfrac{1}{2}Q^TQ + \tfrac{1}{2}Q^TQ\right) d\theta \times \ldots$$

$$= \exp\left(\tfrac{1}{2}Q^TQ\right) \int_b N\left(b \mid 0, \sigma^2\right) \exp\left(\textstyle\sum_t \lambda_t y_t b\right) db \times \exp\left(-\textstyle\sum_t \lambda_t\right)$$

$$= \exp\left(\tfrac{1}{2}Q^TQ\right) \exp\left[\tfrac{1}{2}\sigma^2\left(\textstyle\sum_t \lambda_t y_t\right)^2\right] \exp\left(-\textstyle\sum_t \lambda_t\right)$$

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

$$J(\lambda) = -\log Z(\lambda) = -\tfrac{1}{2}Q^TQ - \tfrac{1}{2}\sigma^2\left(\textstyle\sum_t \lambda_t y_t\right)^2 + \textstyle\sum_t \lambda_t$$

$$= \sum_t \lambda_t - \tfrac{1}{2}\sigma^2\left(\textstyle\sum_t \lambda_t y_t\right)^2 - \tfrac{1}{2}\sum_{t,t'} \lambda_t \lambda_{t'} y_t y_{t'} \left(X_t^T X_{t'}\right)$$

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

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

SVM AGAIN! **WHY +VE?**

MED: Nonlinear Classifiers?

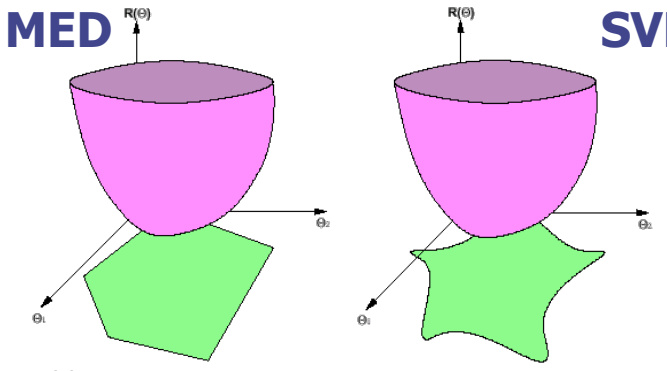
- We can now consider nonlinear discriminant functions *without* using the kernel trick...
- For example a generative-style classifier discriminant function from the ratio of two probability models:

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

Full Gaussians
Tree Structures
Bernoulli
Multinomials
Exponentials
HMMs
Mixtures

- Entropy approach stays convex

$$\int P\left(\Theta\right) y_t f\left(X_t;\Theta\right) \geq 0$$



$$y_t f\left(X_t;\Theta\right) \geq 0$$

SVM breaks

MED: E-Family Ratio Classifiers

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

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

- Proof: Partition function analytically computable

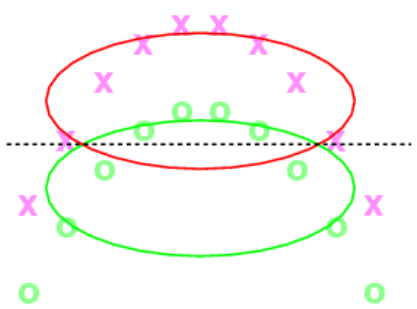
$$Z_{\Theta}=\int P_{0}\left(\Theta\right)\exp\left(\sum_{t}\lambda_{t}y_{t}f\left(X_{t};\Theta\right)\right)d\Theta$$
$$Z_{\Theta}=\int P_{0}\left(\theta_{+}\right)P_{0}\left(\theta_{-}\right)P_{0}\left(b\right)\exp\left(\sum_{t}\lambda_{t}y_{t}\left[\log\frac{P\left(X|\theta_{+}\right)}{P\left(X|\theta_{-}\right)}+b\right]\right)d\Theta$$

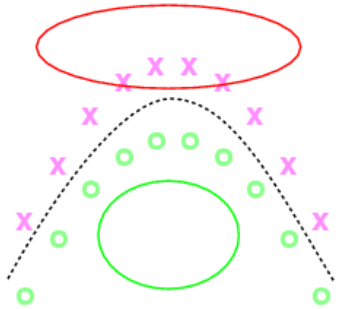
Again Use Non-Informative Prior on b

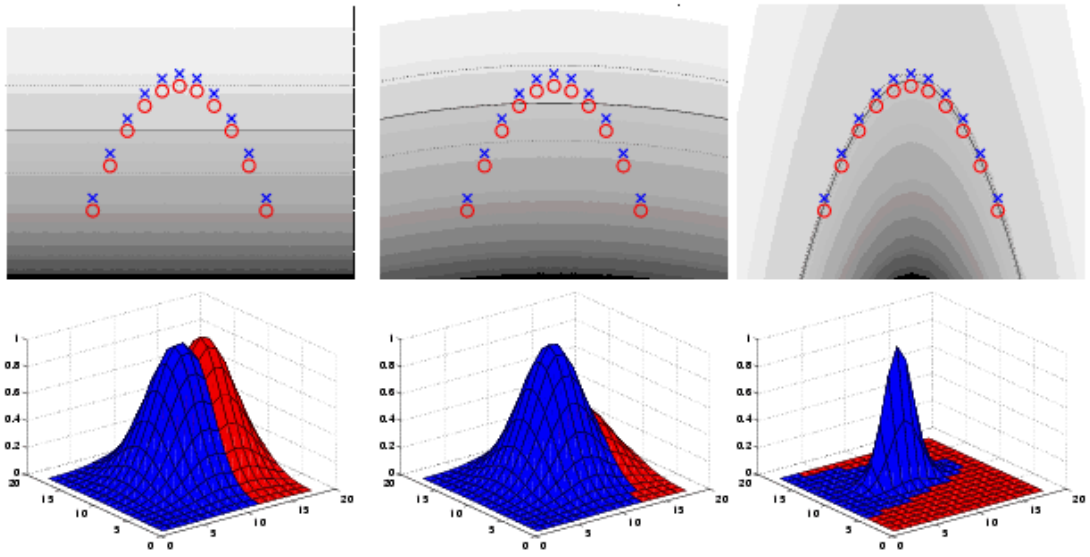
$$Z_{\theta_{\pm}}=\int\exp\left(\tilde{A}\left(\theta_{\pm}\right)+\theta_{\pm}^{T}\chi-\tilde{K}\left(\chi\right)\right)\exp\left(\sum_{t}\lambda_{t}y_{t}\left(A\left(X_{t}\right)+X_{t}^{T}\theta_{\pm}-K\left(\theta_{\pm}\right)\right)\right)d\theta_{\pm}$$
$$Z_{\theta_{\pm}}=\exp\left(-\tilde{K}\left(\chi\right)+\sum_{t}\lambda_{t}y_{t}A\left(X_{t}\right)\right)\times\int\exp\left(\tilde{A}\left(\theta_{\pm}\right)+\theta_{\pm}^{T}\left(\chi+\sum_{t}\lambda_{t}y_{t}X_{t}\right)\right)d\theta_{\pm}$$
$$Z_{\theta_{\pm}}=\exp\left(-\tilde{K}\left(\chi\right)+\sum_{t}\lambda_{t}y_{t}A\left(X_{t}\right)\right)\times\exp\left(\tilde{K}\left(\chi+\sum_{t}\lambda_{t}y_{t}X_{t}\right)\right)$$

MED: Ratio of Gaussians

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







(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