Maximum Entropy Discrimination

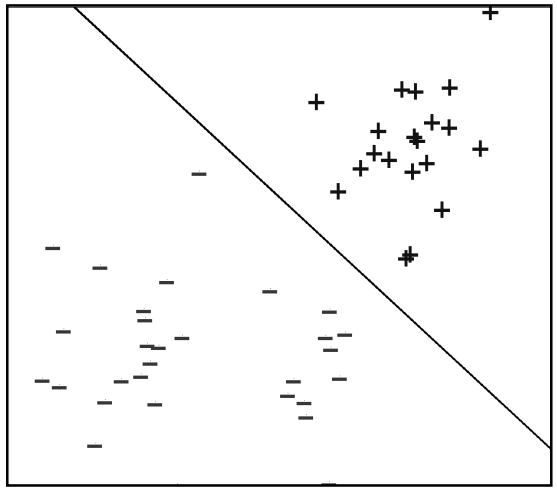
Tommi Jaakkola
MIT

Marina Meila
CMU

Tony Jebara
MIT

Classification

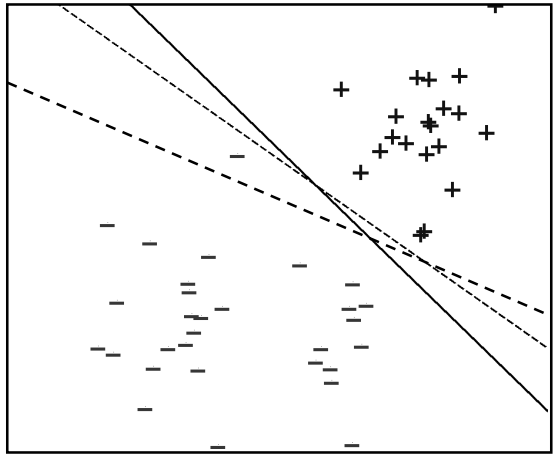
- inputs x, class y = +1, -1
- data D = { (x_1,y_1), ... (x_T,y_T) }



- learn $f^{opt}(x)$ discriminant function from F = {f} family of discriminants
- classify y = sign $f^{opt}(x)$

Model averaging

- many f with near optimal performance



- Instead of <u>choosing</u> f^{opt},
  <u>average</u> over all f in F

$Q(f)$ = weight of f

$$y(x) = sign \int_F Q(f) f(x)$$

$$= sign \langle f(x) \rangle_Q$$

- To specify:
  $F = \{ f \}$ family of discriminant functions
- To learn $Q(f)$ distribution over F

Goal of this work

- Define a discriminative criterion for averaging over models

Advantages

- can incorporate prior
- can use generative model
- computationally feasible
- generalizes to other discrimination tasks

Maximum Entropy Discrimination

given data set $\mathbf{D} = \{ (x_1,y_1), \ldots (x_T,y_T) \}$ find

$$Q_{ME} = \operatorname{argmax}_Q H(Q)$$

s.t. $\quad y_t \langle f(x_t) \rangle_Q \geqslant \gamma \qquad$ for all $t = 1,\ldots,T \qquad (C)$

and some $\gamma > 0$

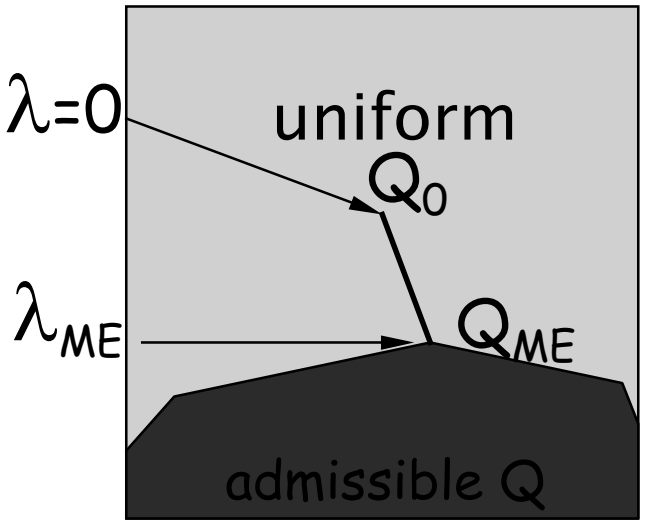
= solution Q_{ME} correctly classifies $\mathbf{D}$

- among all admissible Q, Q_{ME} has max entropy
- max entropy ➡ least specific about f

Solution: Q_{ME} as a projection

- convex problem: Q_{ME} unique
- solution

$$Q_{ME}(f) \sim \exp\{\sum_{t=1}^{T} \lambda_t y_t f(x_t)\}$$



- $\lambda_t \geqslant 0$ Lagrange multipliers
- finding Q_{ME} : start with λ=0 and follow gradient of unsatisfied constraints

Finding the solution

- needed λ_t, $t = 1,\ldots T$
- by solving the dual problem

$$\max_{\lambda} J(\lambda) = \max_{\lambda} [\, - \log Z_+ - \log Z_- - \gamma \textstyle\sum \lambda_t \,]$$

$$\text{s.t. } \lambda_t >= 0 \quad \text{for } t = 1,\ldots T$$

Algorithm

- start with $\lambda_t = 0$ (uniform distribution)
- iterative ascent on $J(\lambda)$ until convergence
- derivative $\partial J / \partial \lambda_t = y_t \langle \log \frac{P^+(x)}{P^-(x)} + b \rangle_{Q(P)} - \gamma$

Q_{ME} as sparse solution

- Classification rule

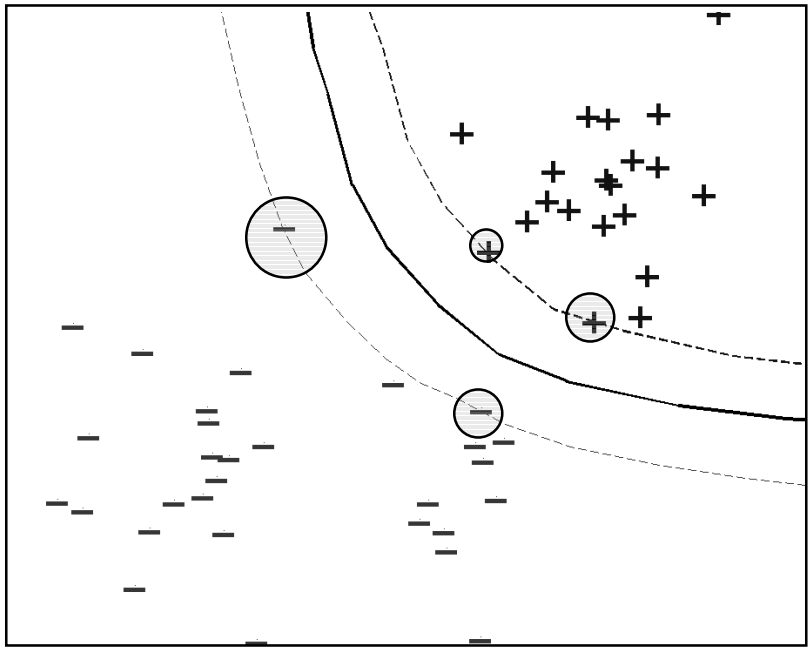
  $y(x) = sign\langle f(x) \rangle_{Q_{ME}}$

- γ is classification margin
- $\lambda_t > 0$ for $y_t \langle f(x_t) \rangle_Q = \gamma$

  χ

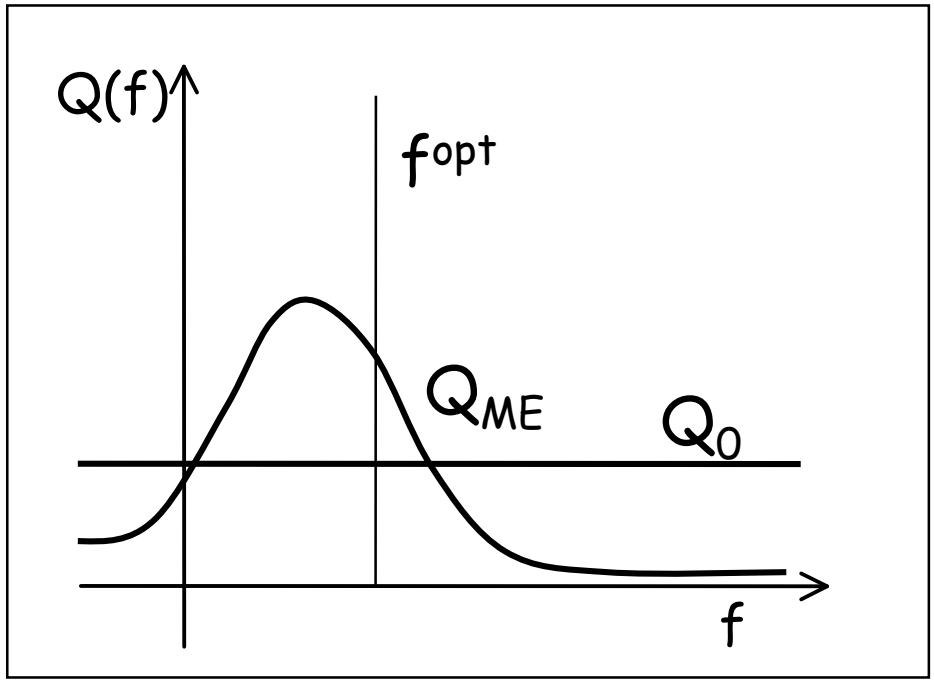
  x_t on the margin

  (support vector!)



Q_{ME} as regularization

- Uniform distribution $Q_0 \Longleftrightarrow \lambda=0$
- "smoothness" of Q = H(Q)
- Q_{ME} is smoothest admissible distribution



Goal of this work

- Define a discriminative criterion for averaging over models ✓

Extensions

⇨ • incorporate prior

⇨ • relationship to support vectors

⇨ • use generative models

⇨ • generalizes to other discrimination tasks

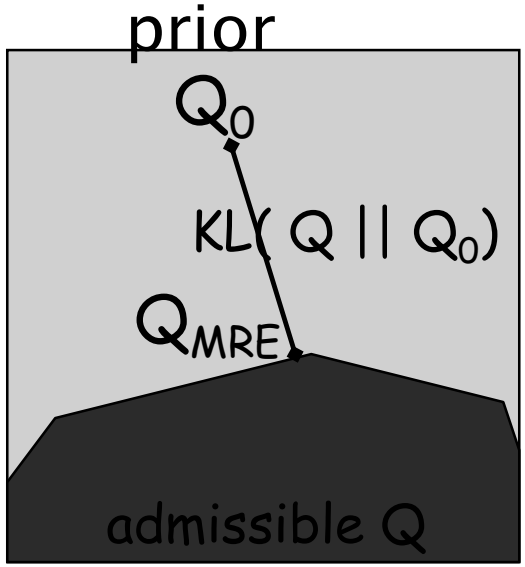
Priors

- prior $Q_0(f)$
- Minimum Relative Entropy Discrimination

$$Q_{MRE} = \text{argmin}_Q \, KL(Q \,||\, Q_0)$$

$$\text{s.t.} \quad y_t < f(x_t) >_Q \geqslant \gamma$$

$$\text{for all} \quad t = 1,\ldots,T \quad (C)$$

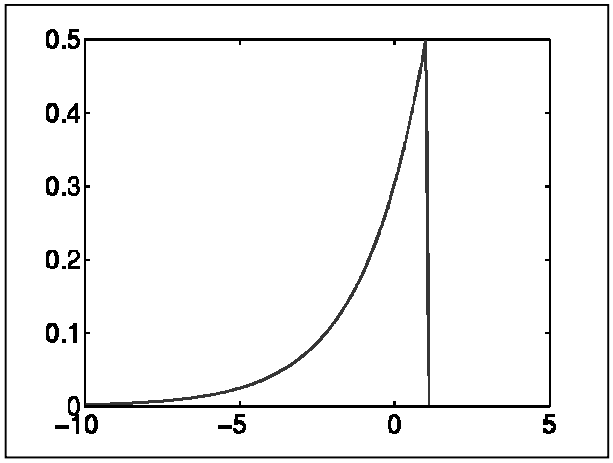


- prior on γ ➡ learn $Q_{MRE}(f, \gamma)$ ➡ soft margin

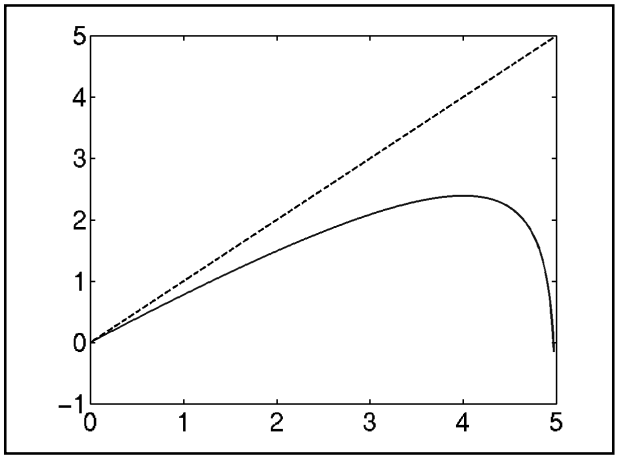
Soft margins

- average also over margin γ
- define $Q_0(f,\gamma) = Q_0(f)\, Q_0(\gamma)$
- constraints $\langle y_t f(x_t) - \gamma \rangle_{Q(f,\gamma)} \geqslant 0$
- learn $Q_{MRE}(f, \gamma) = Q_{MRE}(f)\, Q_{MRE}(\gamma)$

$Q_0(\gamma) = c \exp[c(\gamma-1)]$



Potential as function of λ



Examples: support vector machines

- Theorem

  For $f(x) = \theta \cdot x + b$, $Q_0(\theta) = \text{Normal}(0, I)$, $Q_0(b) =$ non-informative prior, the Lagrange multipliers λ are obtained by maximizing

  $J(\lambda)$ subject to $0 \leqslant \lambda_t \leqslant 0$ and $\sum_t \lambda_t y_t = 0$, where

$$J(\lambda) = \sum_t [\, \lambda_t + \log(1 - \lambda_t/c)\,] - 1/2 \sum_{t,s} \lambda_t \lambda_s y_t y_s x_t . x_s$$

- Separable **D** ⟹ SVM recovered exactly
- Inseparable **D** ⟹ SVM recovered with different misclassification penalty
- Adaptive kernel SVM....

SVM extensions

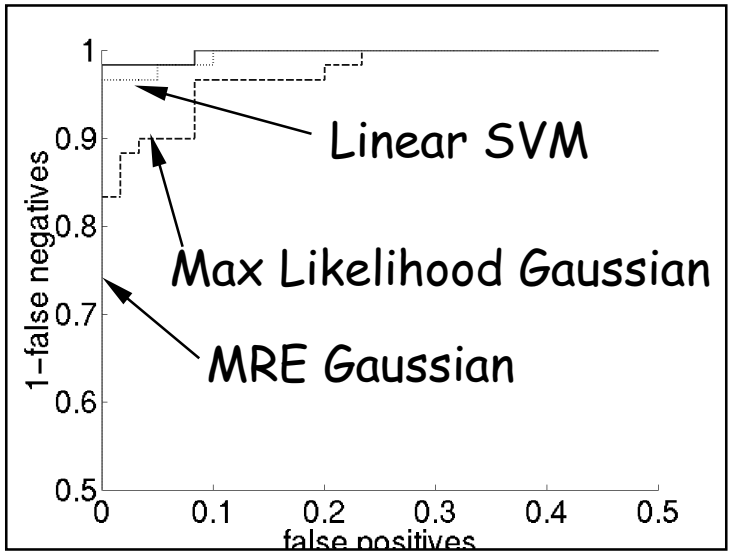
- Example: Leptograpsus Crabs (5 inputs, $T_{train}=80$, $T_{test}=120$)

$$f(x) = \log \frac{P^+(x)}{P^-(x)} + b$$

with $P^+(x) = \text{normal}(x ; m_+, V_+)$

⇒ quadratic classifier

$Q(V_+, V_-)$ = distribution of kernel width



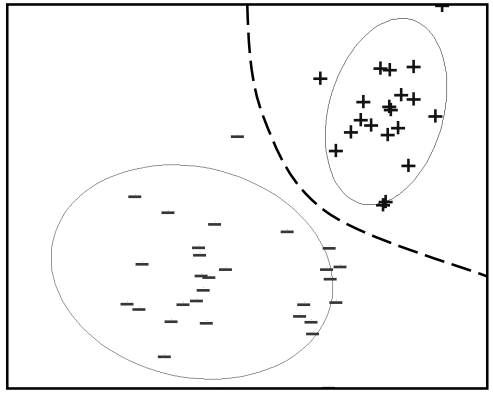
Using generative models

- generative models

  $P^+(x), P^-(x)$ for $y = +1, -1$

- $f(x) = \log \frac{P^+(x)}{P^-(x)} + b$

- learn $Q_{MRE}(P^+, P^-, b, \gamma)$



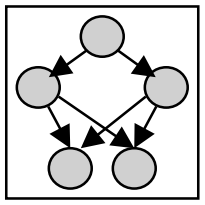
- if $Q_0(P^+, P^- b, \gamma) = Q_0(P^+)\, Q_0(P^-)\, Q_0(b)\, Q_0(\gamma)$

$\Rightarrow$ $Q_{MRE}(P^+, P^-) = Q_{ME}(P^+)\, Q_{ME}(P^-)\, Q_{MRE}(b)\, Q_{MRE}(\gamma)$

(factored prior $\Rightarrow$ factored posterior)

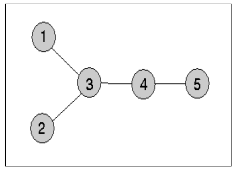
Examples: other distributions

- Multinomial (1 discrete variable) ☑



- Graphical model ☑

(fixed structure, no hidden variables)

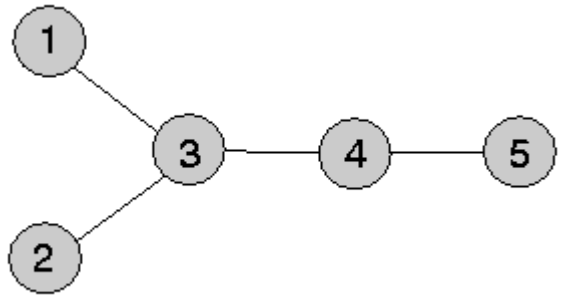


- Tree graphical model ☑

(Q over structures and parameters)

Tree graphical models

- $P(x|\ E, \theta) = P_0(x) \prod_E P_{uv}(x_u x_v|\theta_{uv})$



- prior $Q_0(P) = Q_0(E)\ Q_0(\theta|E)$
  - $Q_0(E) = \prod_E \alpha_{uv}$
  - $Q_0(\theta|E)$ = conjugate prior

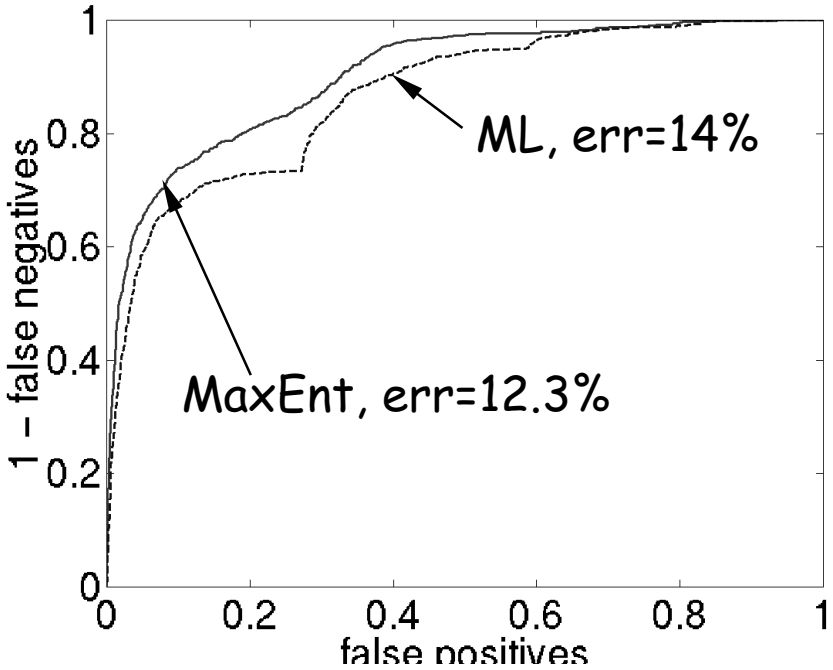
⇒ $Q_0(P)$ conjugate prior over E and θ

⇒ $Q_{MRE}(P) = W_0 \prod_E W_{uv}$

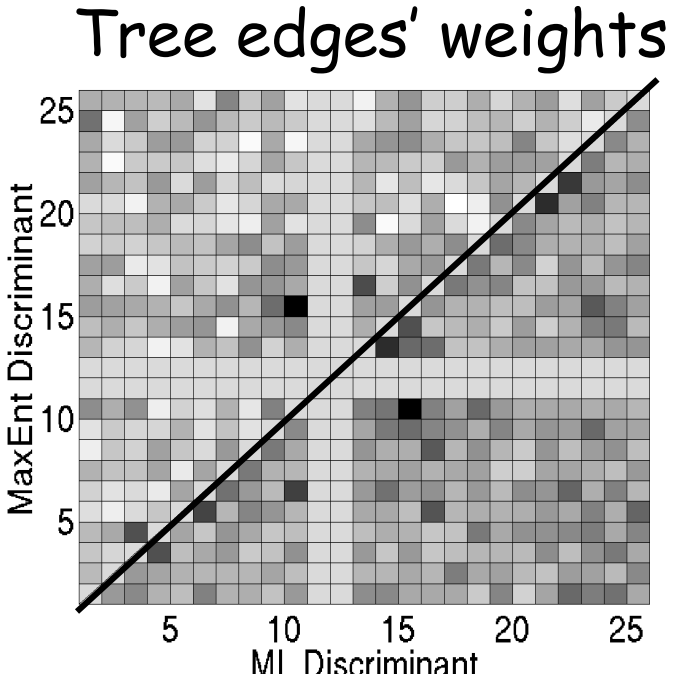
can be integrated analytically

Trees: experiments

- Splice junction classification task
  - 25 inputs, 400 training examples
  - compared with Max Likelihood trees

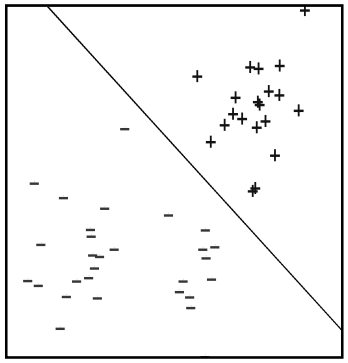


Trees experiments (contd)

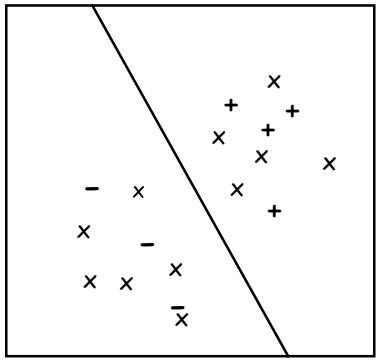


Discrimination tasks

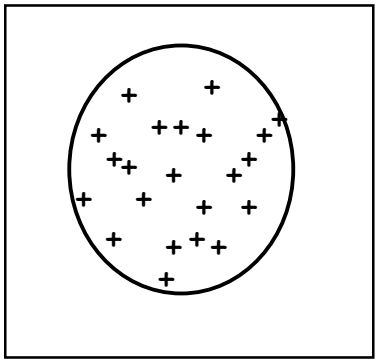
- Classification



- Classification with partially labeled data



- Anomaly detection



Partially labeled data

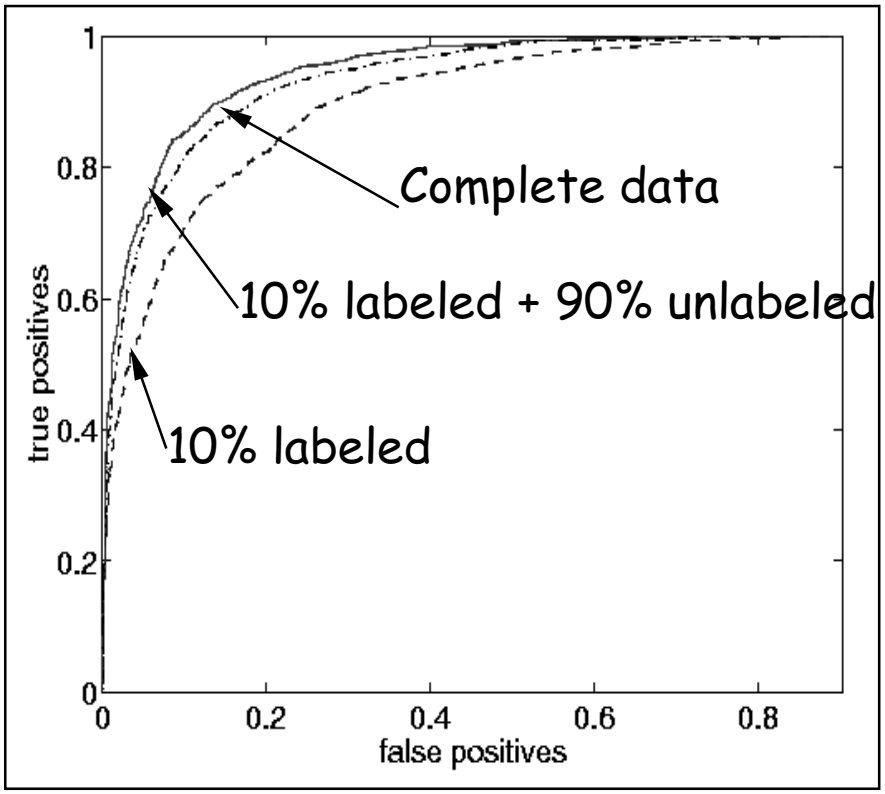
- Problem: given **F** families of discriminants and data set **D** = { $(x_1, y_1) \ldots (x_T, y_T), x_{T+1}, \ldots x_N$ } find

$$Q(f,\gamma,y) = \operatorname{argmin}_Q KL(Q||Q_0)$$

$$\text{s. t. } \langle y_t f(x) - \gamma \rangle_Q \geqslant 0 \quad \text{for all } t = 1,\ldots,T \quad (C)$$

Partially labeled data : experiment

- Splice junction classification
- 25 inputs
- $T_{total}=1000$



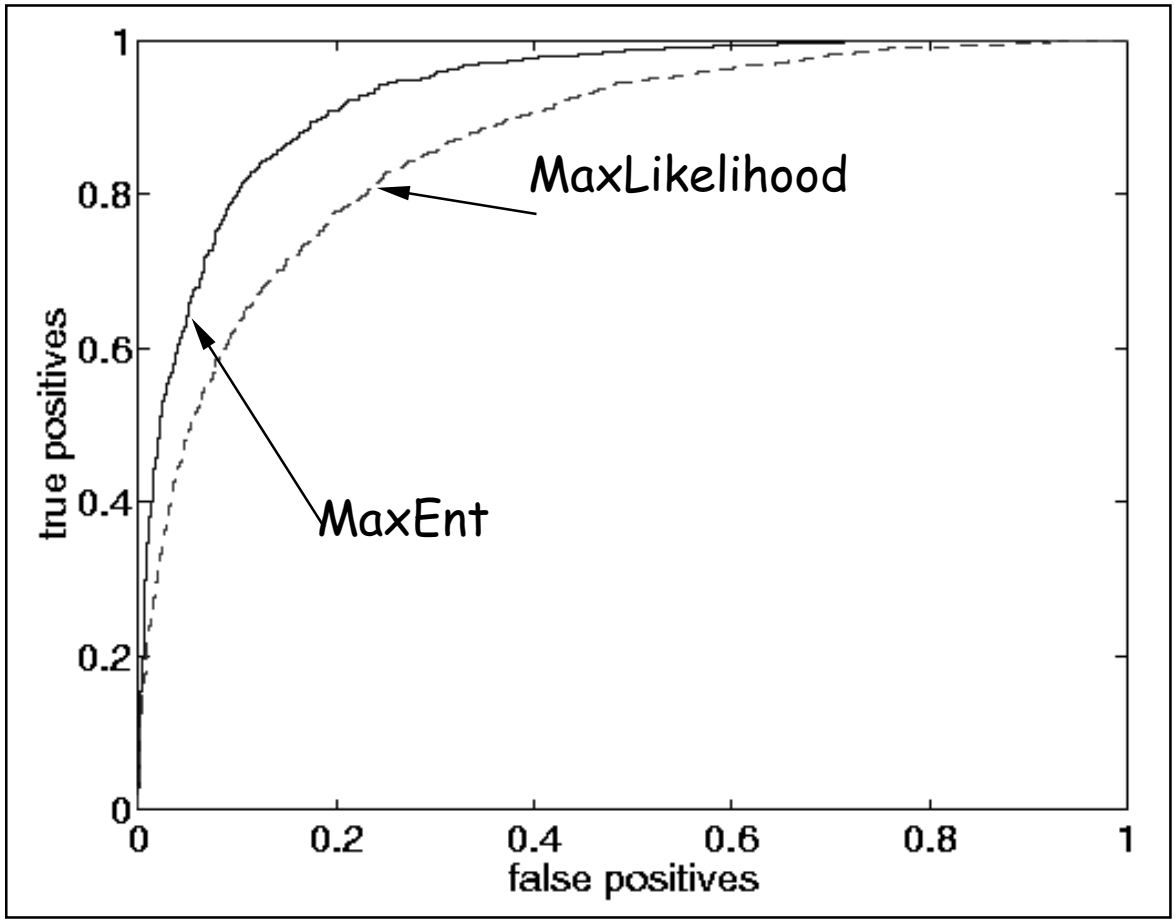
Anomaly detection

- Problem: given **P** = { P } family of generative models and data set **D** = { $x_1, \ldots x_T$ } find Q(P) that

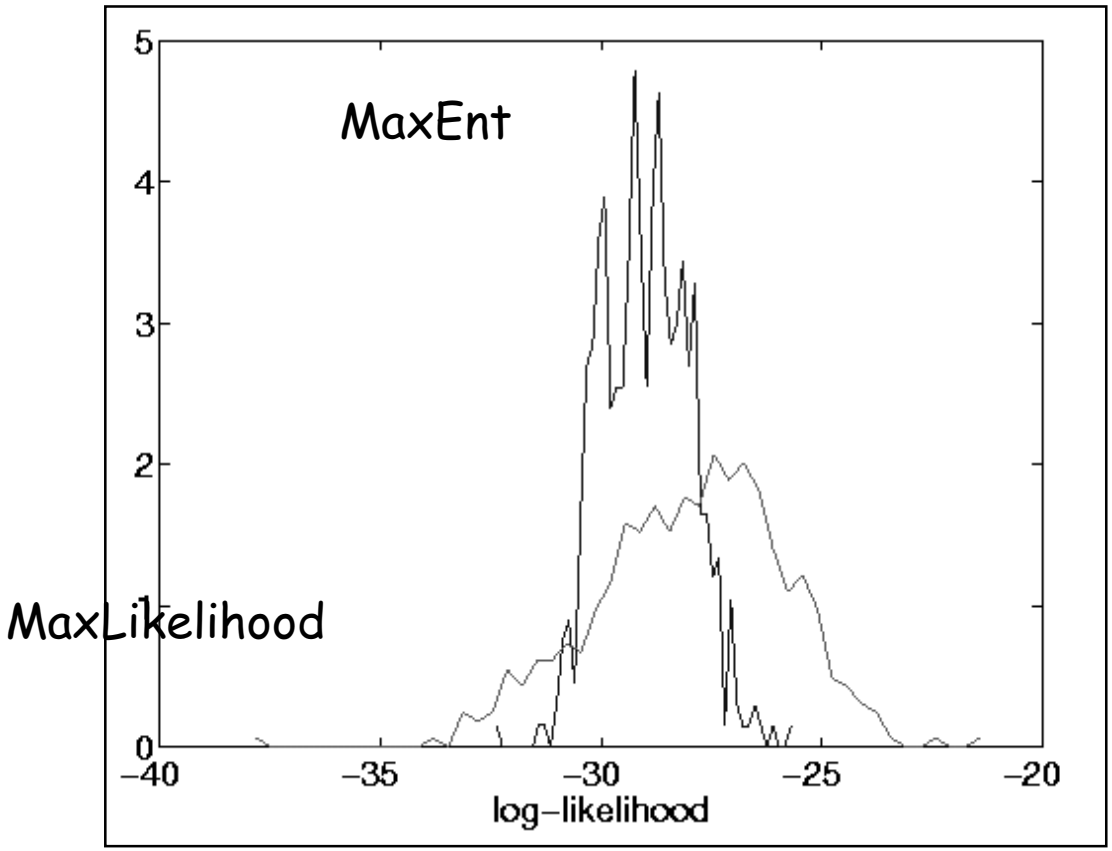
$$Q(P,\gamma) = \operatorname{argmin}_Q KL(Q||Q_0)$$

$$\text{s. t. } \langle \log P(x) - \gamma \rangle_Q \geqslant 0 \qquad \text{for all } t = 1,\ldots,T \qquad (C)$$

Anomaly detection: experiments



Anomaly detection: experiments



Conclusions

- New framework for classification
- Based on regularization in the space of distributions
- Enables use of generative models
- Enables use of priors
- Generalizes to other discrimination tasks