Graphical Models & Kernel Machines

STAT261: Introduction to Machine Learning

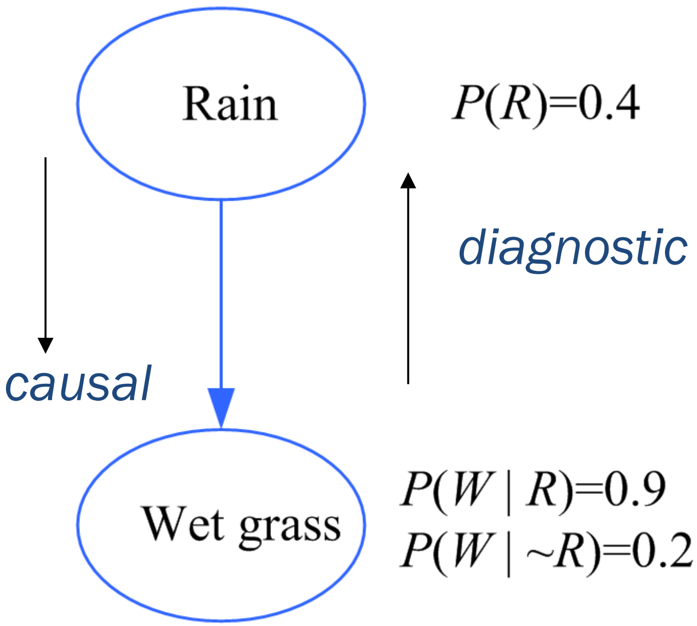
June 1, 2016

Prof. Allie Fletcher

Graphical Models

- Aka Bayesian networks, probabilistic networks
- Nodes are random variables
- Arcs are direct influences between hypotheses
  - Depends in part on which conditional probabilities happen to be known
- The structure is represented as a directed acyclic graph (DAG)
- The parameters are the conditional probabilities in the arcs (Pearl, 1988, 2000; Jensen, 1996; Lauritzen, 1996)

Causes and Bayes' Rule



Diagnostic inference:
Knowing that the grass is wet, what is the probability that rain is the cause?

$$P(R \mid W)=\frac{P(W \mid R)P(R)}{P(W)}$$

$$=\frac{P(W \mid R)P(R)}{P(W \mid R)P(R)+P(W \mid \sim R)P(\sim R)}$$

$$=\frac{0.9 \times 0.4}{0.9 \times 0.4+0.2 \times 0.6}=0.75$$

Conditional Independence

- X and Y are independent if

$$P(X,Y)=P(X)P(Y)$$

- X and Y are conditionally independent given Z if

$$P(X,Y|Z)=P(X|Z)P(Y|Z)$$

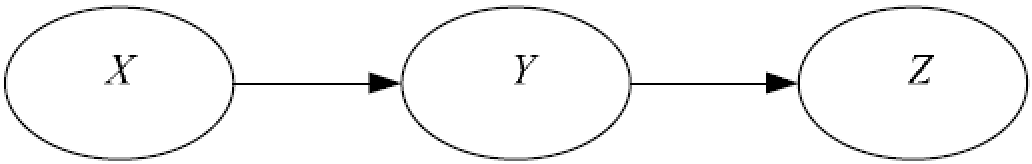
  or

$$P(X|Y,Z)=P(X|Z)$$

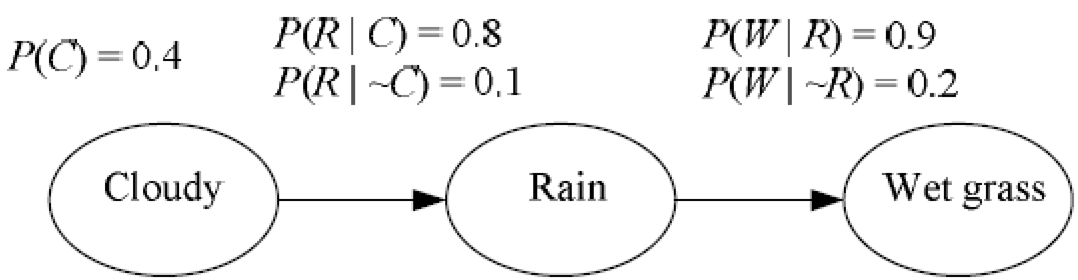
- Three canonical cases: Head-to-tail, Tail-to-tail, head-to-head

Case 1: Head-to-Head

- $P(X,Y,Z)=P(X)P(Y|X)P(Z|Y)$



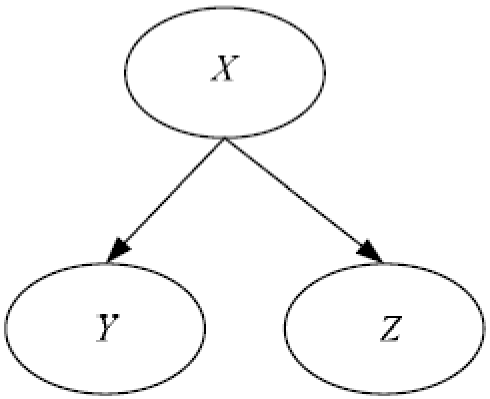
(a) Model

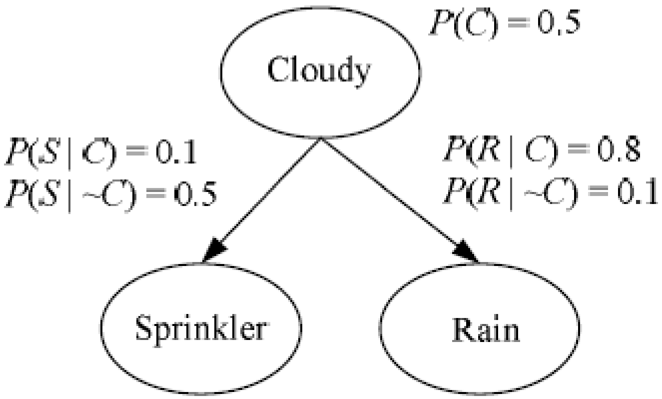


- $P(W|C)=P(W|R)P(R|C)+P(W|\sim R)P(\sim R|C)$

Case 2: Tail-to-Tail

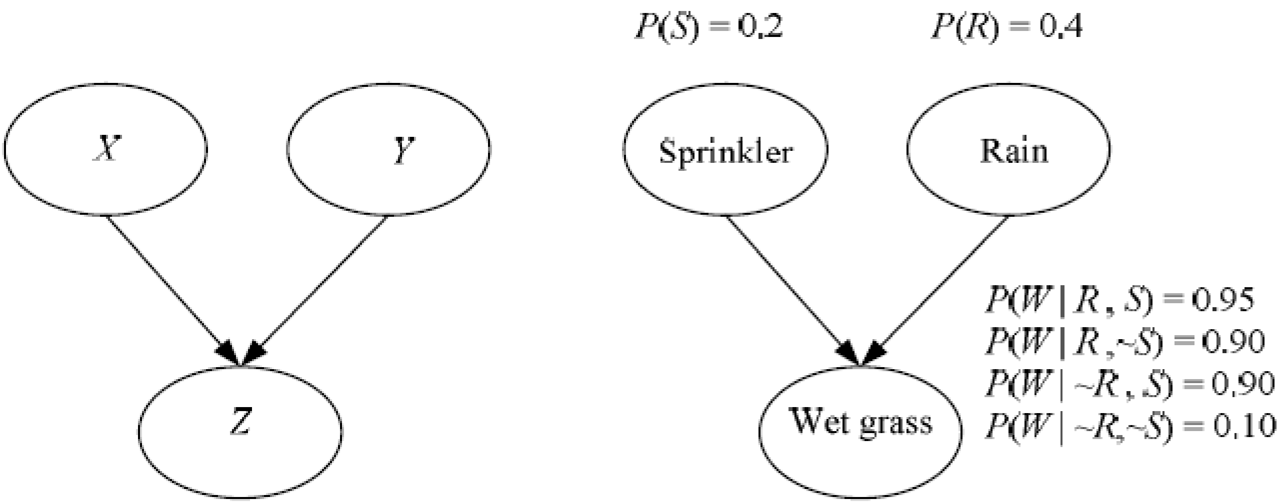
- $P(X,Y,Z)=P(X)P(Y|X)P(Z|X)$



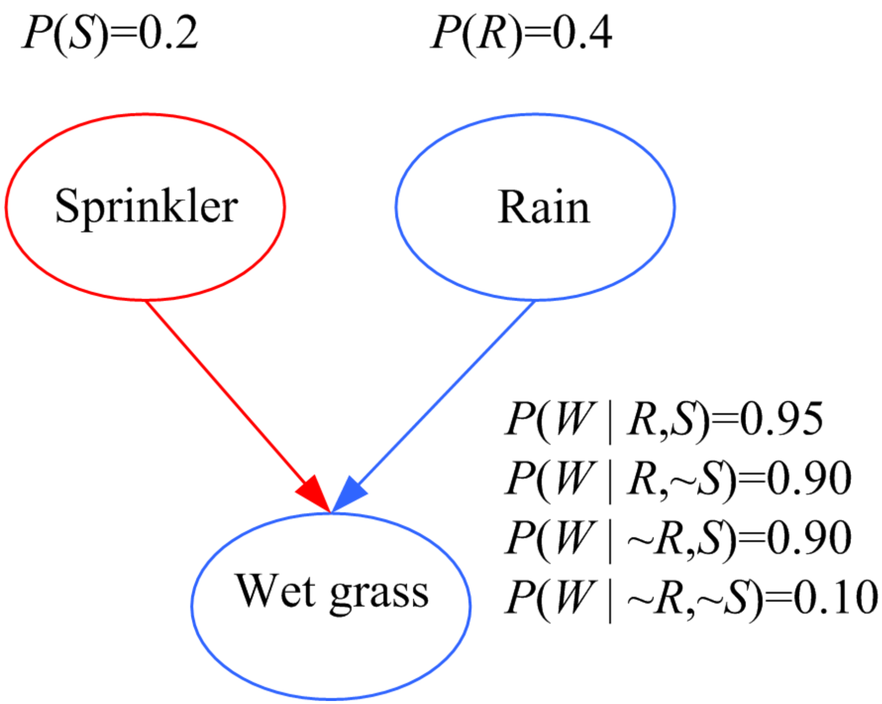


Case 3: Head-to-Head

- $P(X,Y,Z)=P(X)P(Y)P(Z|X,Y)$



Causal vs Diagnostic Inference



Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

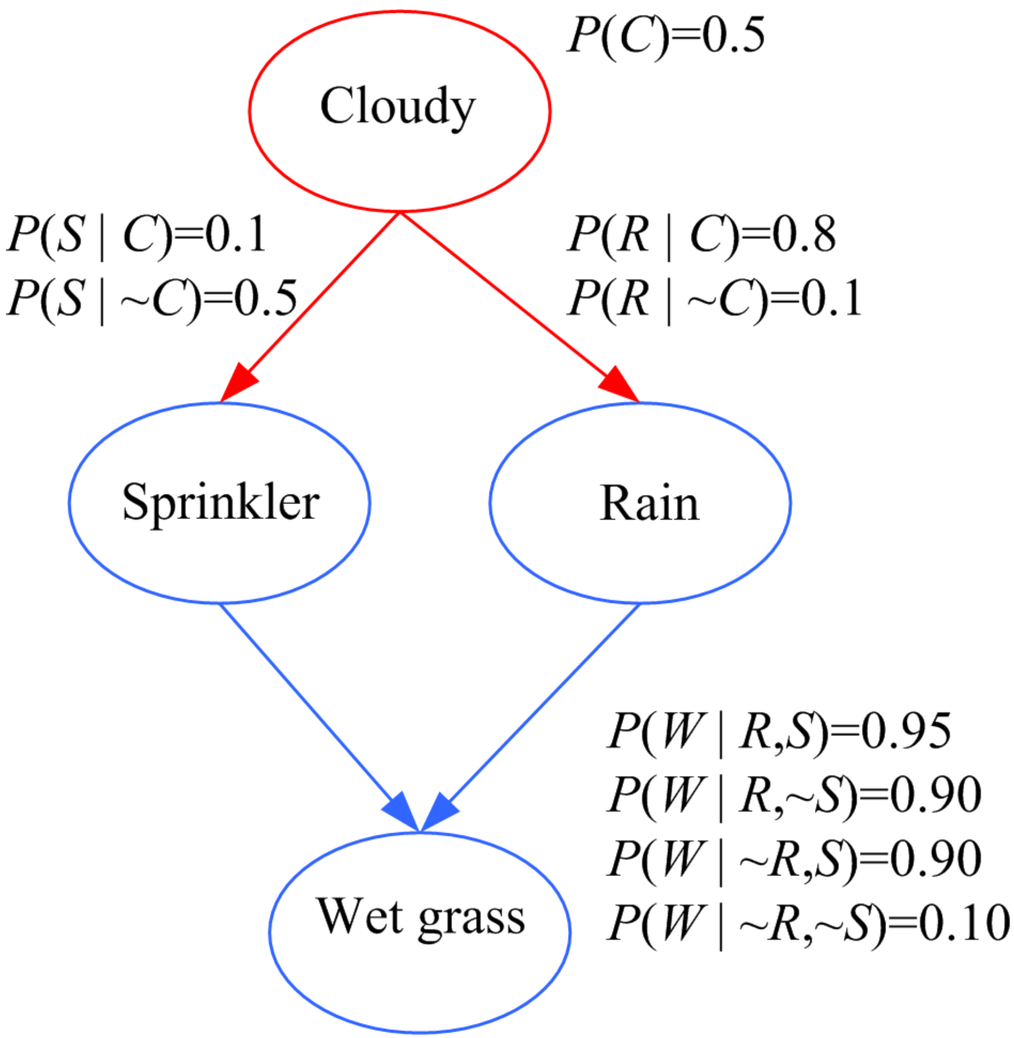
$$
\begin{aligned}
P(W|S) &= P(W|R,S)\, P(R|S) + P(W|\sim R,S)\, P(\sim R|S) \\
&= P(W|R,S)\, P(R) + P(W|\sim R,S)\, P(\sim R) \\
&= 0.95\ 0.4 + 0.9\ 0.6 = 0.92
\end{aligned}
$$

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on? $P(S|W) = 0.35 > 0.2\ P(S)$

$P(S|R,W) = 0.21$

Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.

Causes



Causal inference:

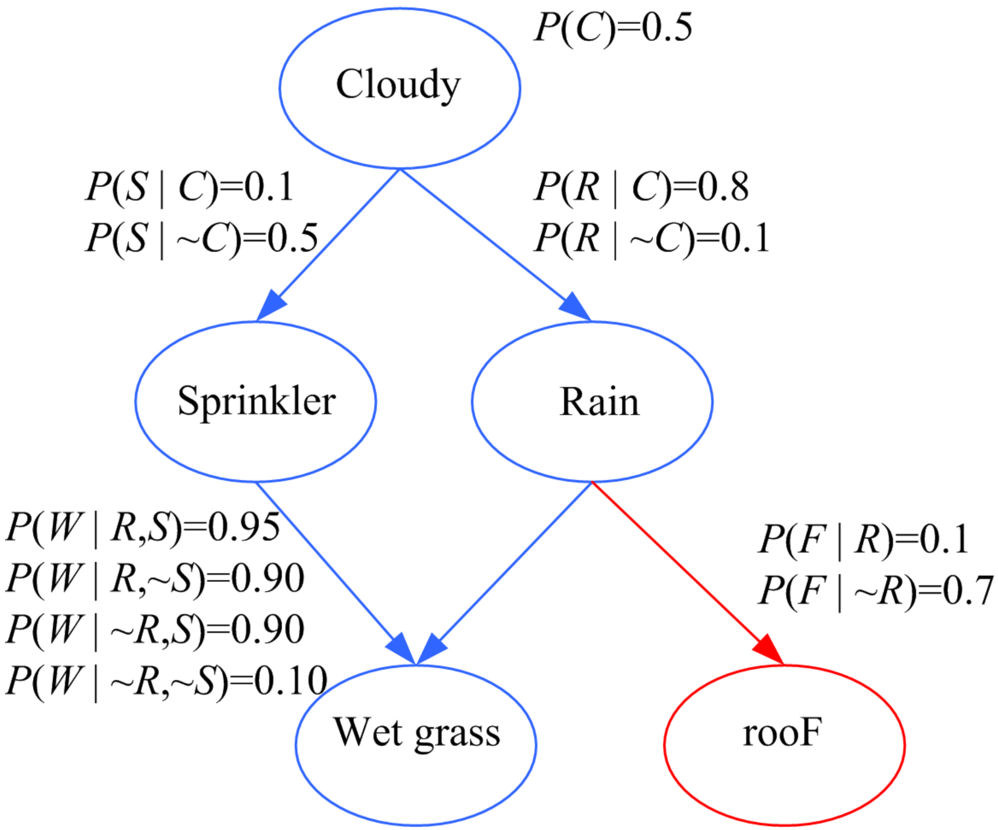
$P(W|C) = P(W|R,S)\,P(R,S|C) +$
$\quad P(W|{\sim}R,S)\,P({\sim}R,S|C) +$
$\quad P(W|R,{\sim}S)\,P(R,{\sim}S|C) +$
$\quad P(W|{\sim}R,{\sim}S)\,P({\sim}R,{\sim}S|C)$

and use the fact that

$P(R,S|C) = P(R|C)\,P(S|C)$

Diagnostic: $P(C|W) = ?$

Exploiting the Local Structure

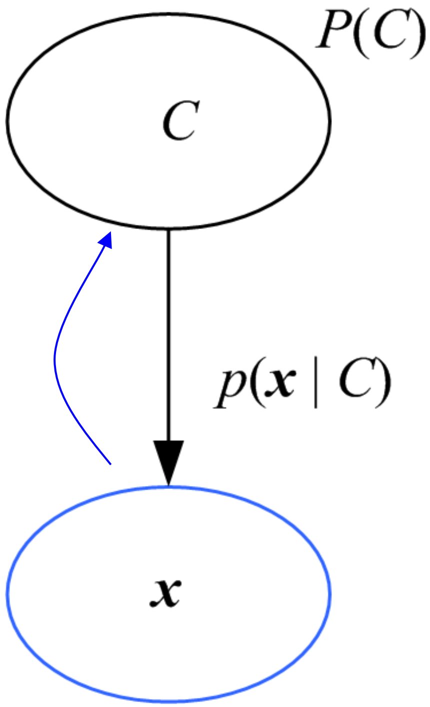


$P(F \mid C) = ?$

$$P(C,S,R,W,F) = P(C)P(S \mid C)P(R \mid C)P(W \mid S,R)P(F \mid R)$$

$$P(x_1, \ldots x_d) = \prod_{i=1}^{d} P(x_i \mid \text{parents}(x_i))$$

Classification



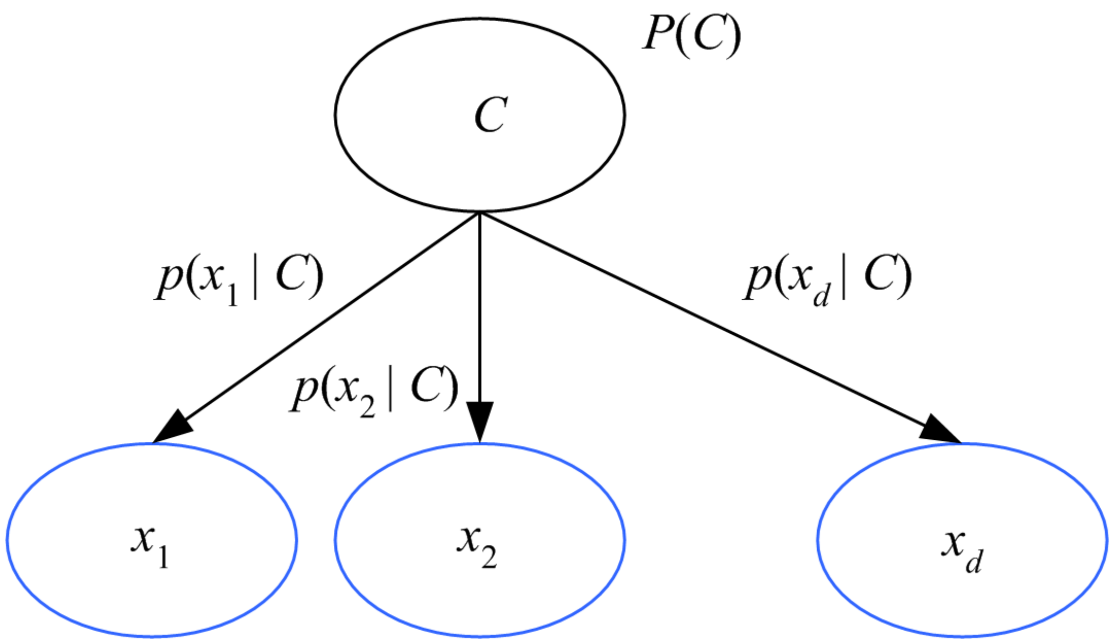
diagnostic

P (C|x)

Bayes' rule inverts the arc:

$$P(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C)P(C)}{p(\mathbf{x})}$$

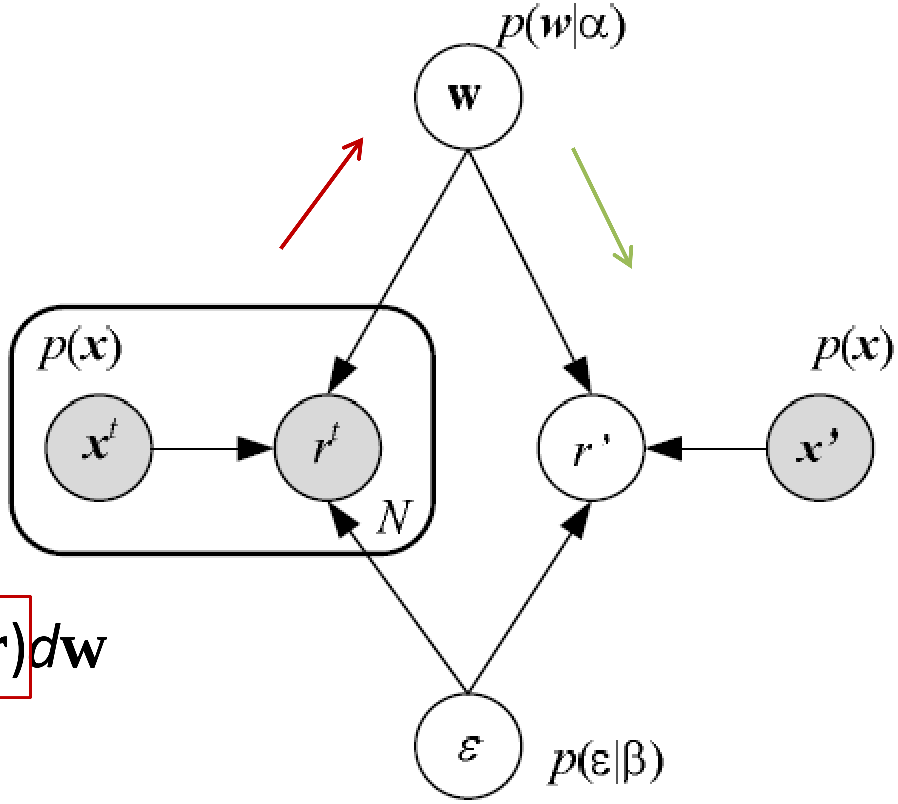
Naive Bayes' Classifier



Given C, x_j are independent:

$$p(\boldsymbol{x}|C) = p(x_1|C)\, p(x_2|C) \ldots p(x_d|C)$$

Linear Regression



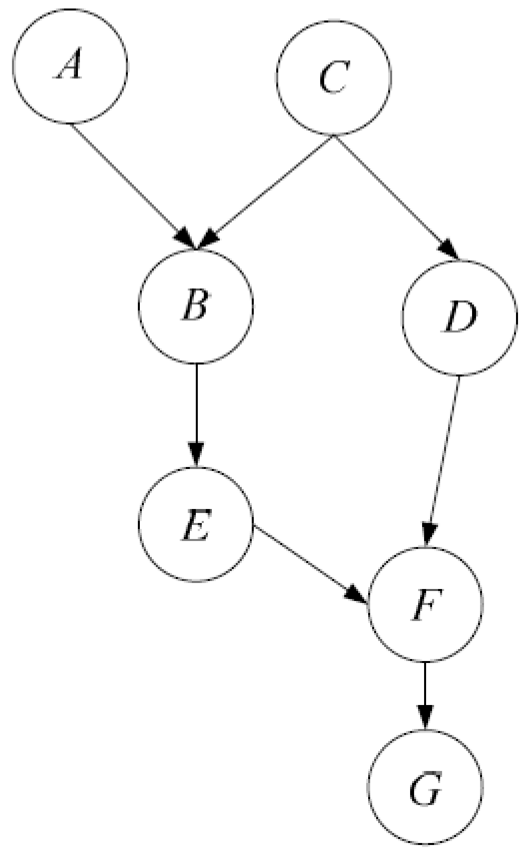
$$p(r'|\mathbf{x}',\mathbf{r},\mathbf{X}) = \int \boxed{p(r'|\mathbf{x}',\mathbf{w})}\boxed{p(\mathbf{w}|\mathbf{X},\mathbf{r})}d\mathbf{w}$$

$$= \int p(r'|\mathbf{x}',\mathbf{w})\frac{p(\mathbf{r}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{r})}d\mathbf{w}$$

$$\propto \int p(r'|\mathbf{x}',\mathbf{w})\prod_t p(r^t|\mathbf{x}^t,\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

d-Separation

- A path from node *A* to node *B* is blocked if
  a) The directions of edges on the path meet head-to-tail (case 1) or tail-to-tail (case 2) and the node is in *C*, or
  b) The directions of edges meet head-to-head (case 3) and neither that node nor any of its descendants is in *C*.
- If all paths are blocked, *A* and *B* are d-separated (conditionally independent) given *C*.



BCDF is blocked given *C*.
BEFG is blocked by *F*.
BEFD is blocked unless *F* (or *G*) is given.

Kernel Machines

- Discriminant-based: No need to estimate densities first
- Define the discriminant in terms of support vectors
- The use of kernel functions, application-specific measures of similarity
- No need to represent instances as vectors
- Convex optimization problems with a unique solution

Optimal Separating Hyperplane

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \mathbf{x}^t \in C_1 \\ -1 & \text{if } \mathbf{x}^t \in C_2 \end{cases}$$

find $\mathbf{w}$ and w_0 such that

$$\mathbf{w}^T\mathbf{x}^t + w_0 \geq +1 \text{ for } r^t = +1$$

$$\mathbf{w}^T\mathbf{x}^t + w_0 \leq +1 \text{ for } r^t = -1$$

which can be rewritten as

$$r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1$$

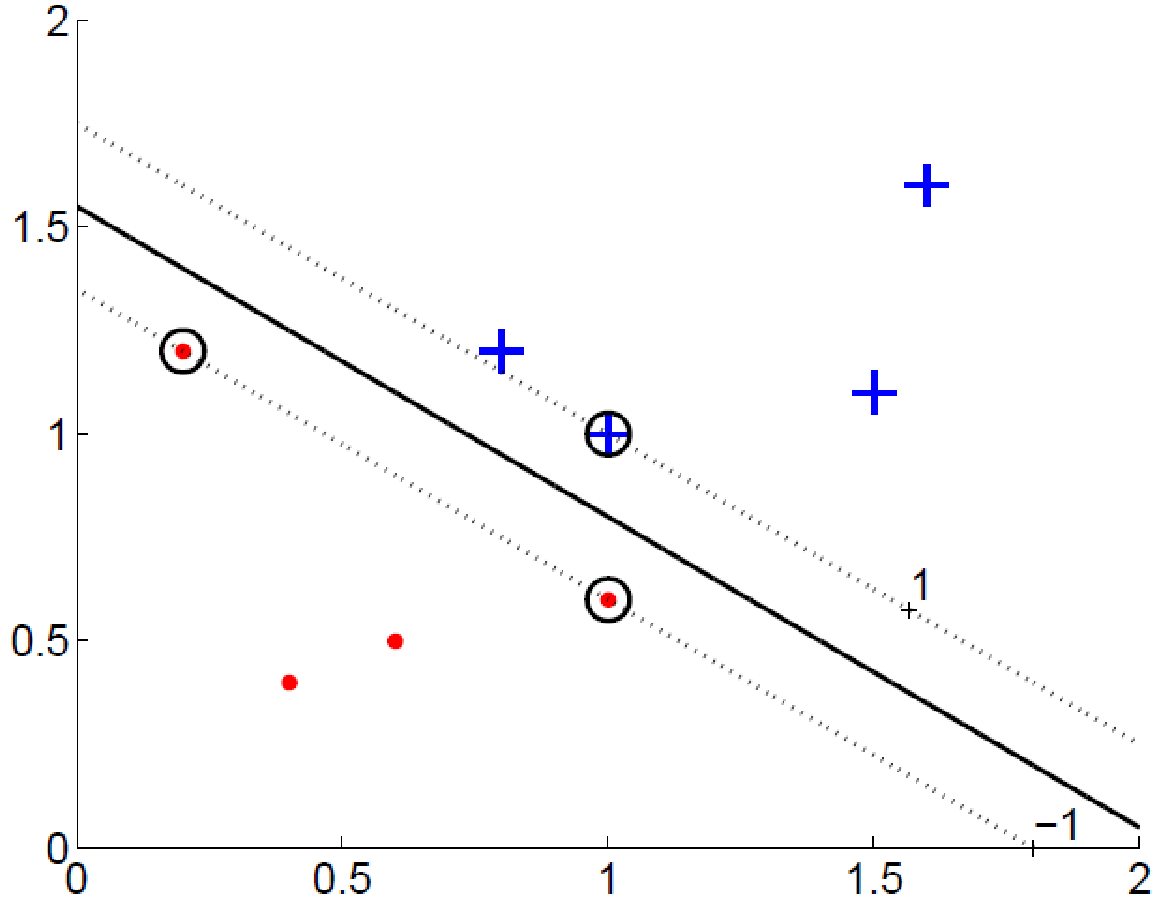
(Cortes and Vapnik, 1995; Vapnik, 1995)

Margin

- Distance from the discriminant to the closest instances on either side
- Distance of x to the hyperplane is $\dfrac{\left|\mathbf{w}^T\mathbf{x}^t + w_0\right|}{\|\mathbf{w}\|}$
- We require $\dfrac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|} \geq \rho, \forall t$
- For a unique sol'n, fix $\rho\|w\|=1$, and to max margin

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

Margin



$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t \left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1\right]$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) + \sum_{t=1}^{N} \alpha^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$

$$
\begin{aligned}
L_d &= \frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) - \mathbf{w}^T\sum_t \alpha^t r^t \mathbf{x}^t - w_0\sum_t \alpha^t r^t + \sum_t \alpha^t \\
&= -\frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) + \sum_t \alpha^t \\
&= -\frac{1}{2}\sum_t\sum_s \alpha^t\alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t
\end{aligned}
$$

subject to $\sum_{\mathrm{t}} \alpha^t r^t = 0$ and $\alpha^t \geq 0, \forall t$

Most α^t are 0 and only a small number have $\alpha^t > 0$; they are the support vectors

Soft Margin Hyperplane

- Not linearly separable

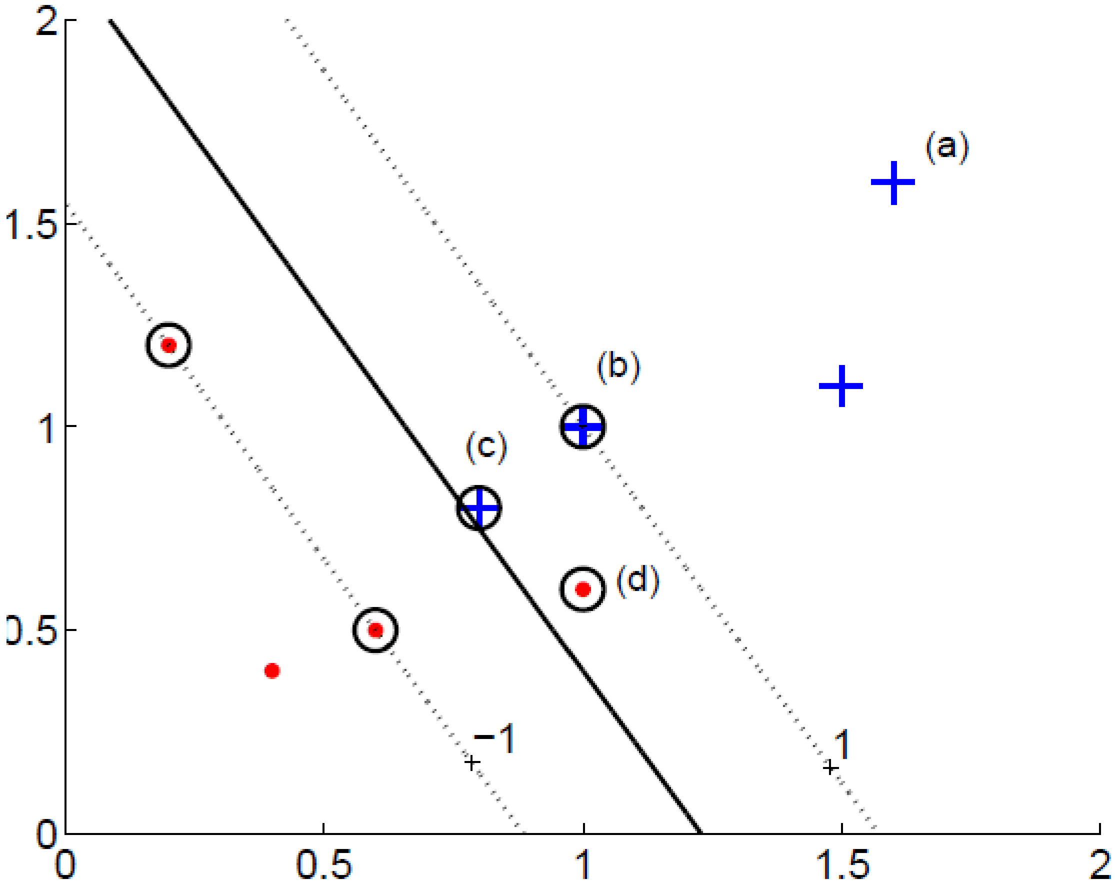
$$r^t\left(\mathbf{w}^T x^t + w_0\right) \geq 1 - \xi^t$$

- Soft error

$$\sum_t \xi^t$$

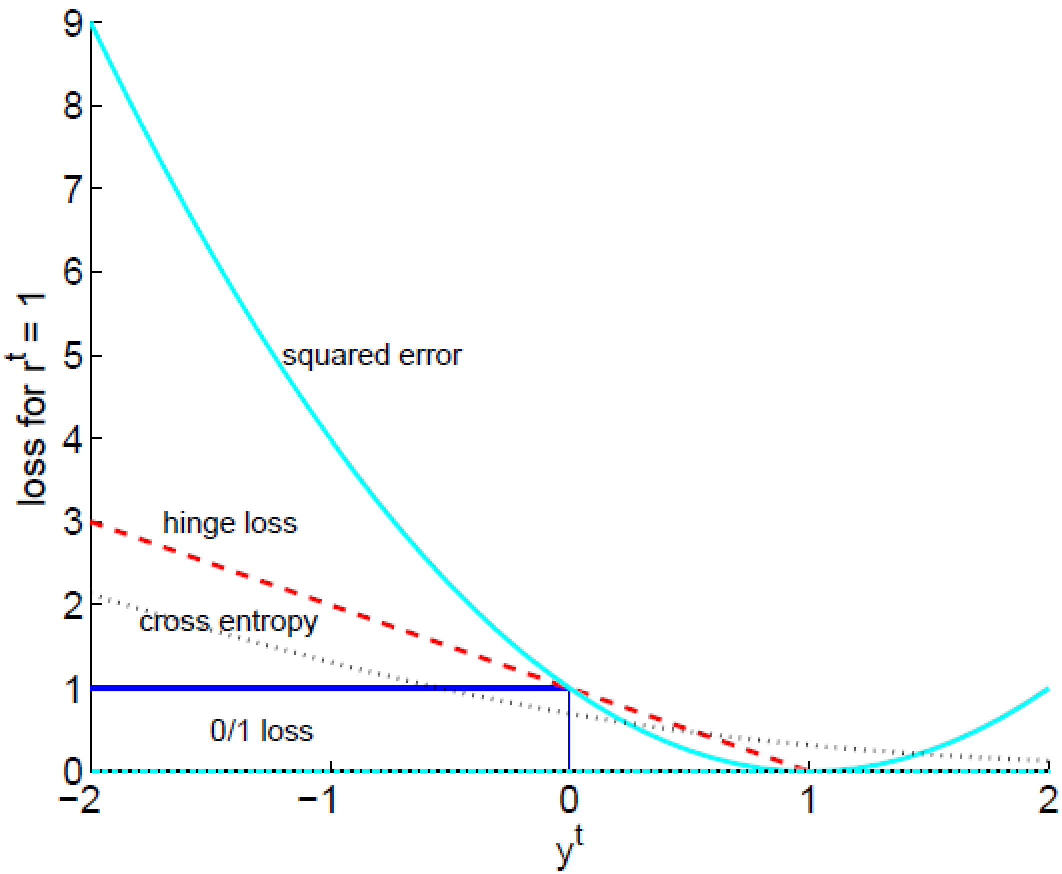
- New primal is

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T x^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$



2
1.5
1
0.5
0
0
0.5
1
1.5
2
(a)
(b)
(c)
(d)
−1
1

Hinge Loss



$$: \begin{cases} 0 & \text{if } y^t r^t \geq 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}$$

ν-SVM

$$\min \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{N}\sum_t \xi^t$$

subject to

$$r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq \rho - \xi^t, \xi^t \geq 0, \rho \geq 0$$

$$L_d = -\frac{1}{2}\sum_{t=1}^{N}\sum_s \alpha^t\alpha^s r^t r^s \left(x^t\right)^T x^s$$

subject to

$$\sum_{\mathrm{t}} \alpha^t r^t = 0, 0 \leq \alpha^t \leq \frac{1}{N}, \sum_t \alpha^t \leq \nu$$

ν controls the fraction of support vectors

UCLA

Kernel Trick

- Preprocess input *x* by basis functions

$$z = \varphi(x) \qquad g(z) = w^T z$$

$$g(x) = w^T \varphi(x)$$

- The SVM solution

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \boldsymbol{\varphi}(\mathbf{x}^t)$$

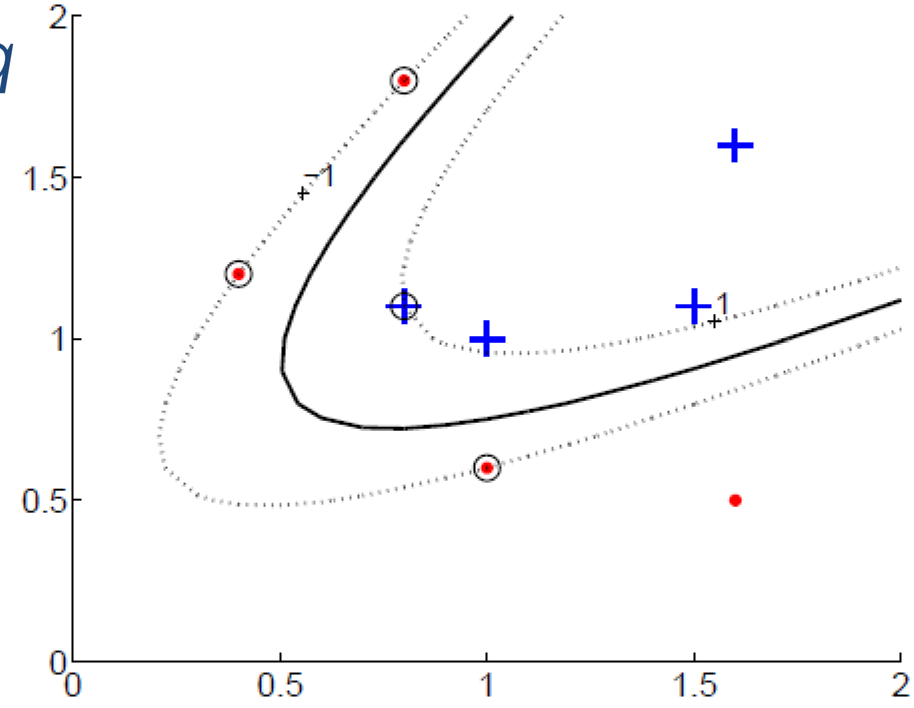
$$g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{\boldsymbol{\varphi}(\mathbf{x}^t)^T \boldsymbol{\varphi}(\mathbf{x})}$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{K(\mathbf{x}^t, \mathbf{x})}$$

Vectorial Kernels

- Polynomials of degree q

$$K\left(\mathbf{x}^t,\mathbf{x}\right)=\left(\mathbf{x}^T\mathbf{x}^t+1\right)^q$$



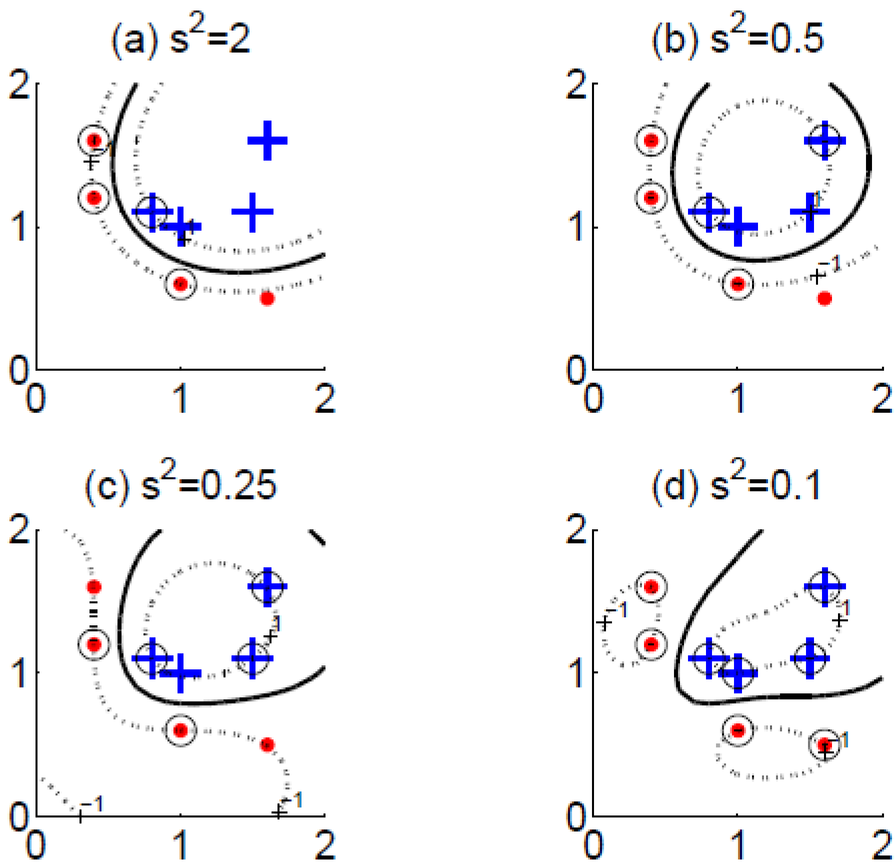
$$\begin{aligned}K(\mathbf{x},\mathbf{y})&=\left(\mathbf{x}^T\mathbf{y}+1\right)^2\\&=\left(x_1y_1+x_2y_2+1\right)^2\\&=1+2x_1y_1+2x_2y_2+2x_1x_2y_1y_2+x_1^2y_1^2+x_2^2y_2^2\end{aligned}$$

$$\phi(\mathbf{x})=\left[1,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_1x_2,x_1^2,x_2^2\right]^T$$

Vectorial Kernels

- Radial-basis functions:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \exp\left[-\frac{\left\|\mathbf{x}^t - \mathbf{x}\right\|^2}{2s^2}\right]$$



Defining kernels

- Kernel "engineering"
- Defining good measures of similarity
- String kernels, graph kernels, image kernels, ...
- Empirical kernel map: Define a set of templates $\boldsymbol{m}_i$ and score function $s(\boldsymbol{x},\boldsymbol{m}_i)$

  $\phi(\boldsymbol{x}^t)=[s(\boldsymbol{x}^t,\boldsymbol{m}_1), s(\boldsymbol{x}^t,\boldsymbol{m}_2),\ldots, s(\boldsymbol{x}^t,\boldsymbol{m}_M)]$

  and

  $K(\boldsymbol{x},\boldsymbol{x}^t)=\phi(\boldsymbol{x})^T \phi(\boldsymbol{x}^t)$

Kernel Machines for Ranking

- We require not only that scores be correct order but at least +1 unit margin.
- Linear case:

$$\min \frac{1}{2}\|\mathbf{w}_i\|^2 + C\sum_t \xi_i^t$$

subject to

$$\mathbf{w}^T\mathbf{x}^u \geq \mathbf{w}^T\mathbf{x}^v + 1 - \xi^t, \forall t : r^u \prec r^v, \xi_i^t \geq 0$$

UCLA

Large Margin Nearest Neighbor

- Learns the matrix **M** of Mahalanobis metric

  $D(\boldsymbol{x}^i, \boldsymbol{x}^j) = (\boldsymbol{x}^i - \boldsymbol{x}^j)^T \mathbf{M} (\boldsymbol{x}^i - \boldsymbol{x}^j)$

- For three instances i, j, and l, where i and j are of the same class and l different, we require

  $D(\boldsymbol{x}^i, \boldsymbol{x}^l) > D(\boldsymbol{x}^i, \boldsymbol{x}^j) + 1$

  and if this is not satisfied, we have a slack for the difference and we learn M to minimize the sum of such slacks over all i,j,l triples (j and l being one of k neighbors of i, over all i)

Learning a Distance Measure

- LMNN algorithm (Weinberger and Saul 2009)

$$(1-\mu)\sum_{i,j}\mathcal{D}(\mathbf{x}^i,\mathbf{x}^j)+\mu\sum_{i,j,l}(1-y_{il})\xi_{ijl}$$

subject to

$$\begin{aligned}\mathcal{D}(\mathbf{x}^i,\mathbf{x}^l) &\geq \mathcal{D}(\mathbf{x}^i,\mathbf{x}^j)+1-\xi^{ijl}, \text{ if } \boldsymbol{r}^i=\boldsymbol{r}^j \text{ and } \boldsymbol{r}^i\neq\boldsymbol{r}^l\\ \xi^{ijl} &\geq 0\end{aligned}$$

- LM ar approach where $\mathbf{M}=\mathbf{L}^T\mathbf{L}$ and learns $\mathbf{L}$

Kernel Dimensionality Reduction

- Kernel PCA does PCA on the kernel matrix (equal to canonical PCA with a linear kernel)
- Kernel LDA, CCA