Linear Discriminant Analysis

- Example: binary (two class) classification

Logit: $$\log\frac{\Pr(G=1|X=x)}{1-\Pr(G=1|X=x)} = \log\frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)} = \beta_0 + x^T\beta$$

- The posterior probability

Q

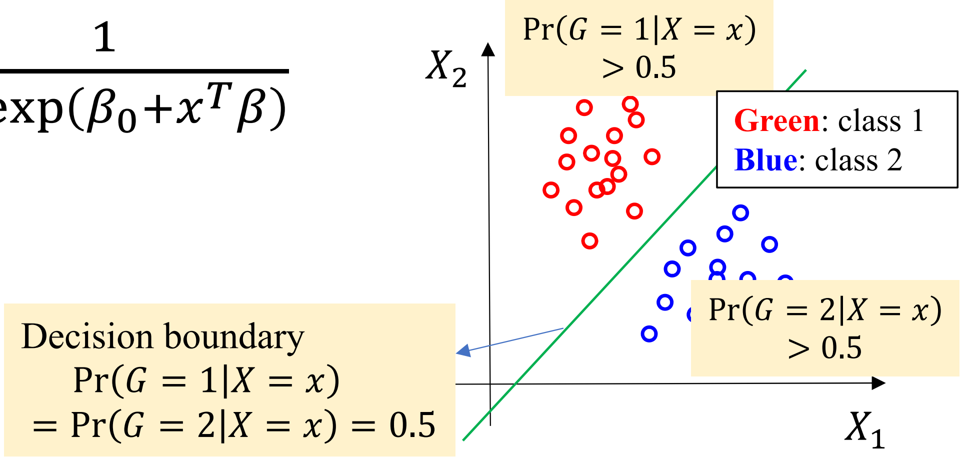
$$\Pr(G=1|X=x) = \frac{\boxed{\exp(\beta_0+x^T\beta)}}{1+\exp(\beta_0+x^T\beta)},$$

$\exp(x) = e^x$

$$\Pr(G=2|X=x) = \frac{1}{1+\exp(\beta_0+x^T\beta)}$$

- Decision boundary

$$\{x|\beta_0 + x^T\beta \ = 0\}$$



$\Pr(G=1|X=x)$ > 0.5

Green: class 1
Blue: class 2

$\Pr(G=2|X=x)$ > 0.5

Decision boundary
$\Pr(G=1|X=x)$
$= \Pr(G=2|X=x) = 0.5$

Fisher's Formulation of Discriminant Analysis

LDA: Approach 1

1. Estimating $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Discriminant function
$$\delta_k(x) = x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k + \log\hat{\pi}_k$$
3. Classify to class k that maximizes the discriminant function
$$\hat{G}(x) = \underset{k\in\mathcal{G}}{\operatorname{argmax}}\,\delta_k(x)$$

Q: why data sphering makes $\hat{\boldsymbol{\Sigma}}^* = \mathbf{I}$?

Hint: $\hat{\boldsymbol{\Sigma}} = \frac{\sum_{k=1}^{K}\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T}{N-K}$

LDA: Approach 2

1. Estimating $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}_k$ and $\hat{\pi}_k$
2. Eigen-decomposition:
$$\hat{\boldsymbol{\Sigma}} = \mathbf{UDU}^T$$
3. Data sphering ($\hat{\Sigma}^* = \mathbf{I}$)
   - ❑ $x^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T x = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}x$
   - ❑ $\hat{\mu}_k^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\hat{\mu}_k = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\hat{\mu}_k$
4. Classify to its closest class centroid in the transformed space
$$G(x) = \underset{k\in\mathcal{G}}{\operatorname{argmin}}\frac{1}{2}\|x^* - \hat{\mu}_k^*\|^2 - \ln\hat{\pi}_k$$

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1-1}\ \theta_2^{\beta_2-1} \ldots \theta_k^{\beta_k-1}}{B(\beta_1, \ldots, \beta_k)} \sim \mathrm{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

and MAP estimate is therefore

$$\hat{\theta}_i^{MAP} = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k}(\alpha_j + \beta_j - 1)}$$

Q

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

$\ell(\theta,\pi) =$ ______

Q $\rightarrow$

$$\underline{\hat{\pi}_k} = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\frac{\partial \ell}{\partial \theta} = 0 \qquad \frac{\partial \ell}{\partial \pi} = 0$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

Only difference:
“imaginary” examples

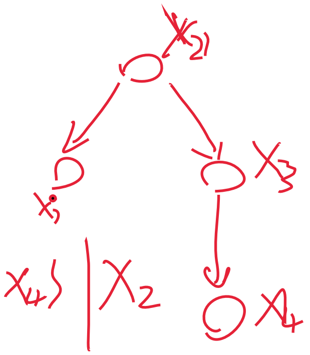
$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

$\ln P(\text{Data}|\theta,\pi)\, P(\pi)$

What is the Bayes Network for Naïve Bayes?

Q

Conditional Independence, Revisited



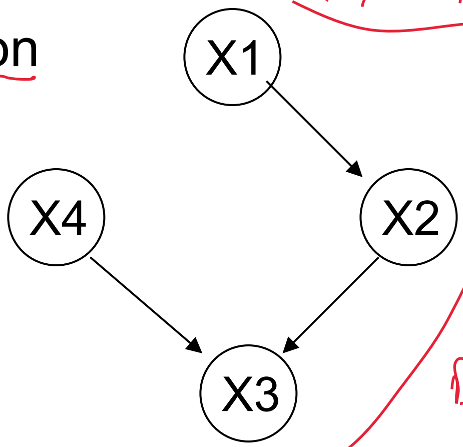
$X_1 \perp\!\!\!\perp \{X_3, X_4\} \mid X_2$

- We said:
  - Each node is conditionally independent of its non-descendents, given its immediate parents.

- Does this rule give us all of the conditional independence relations implied by the Bayes network?
  - No!
  - E.g., X1 and X4 are conditionally indep given {X2, X3}
  - But X1 and X4 not conditionally indep given X3
  - For this, we need to understand D-separation

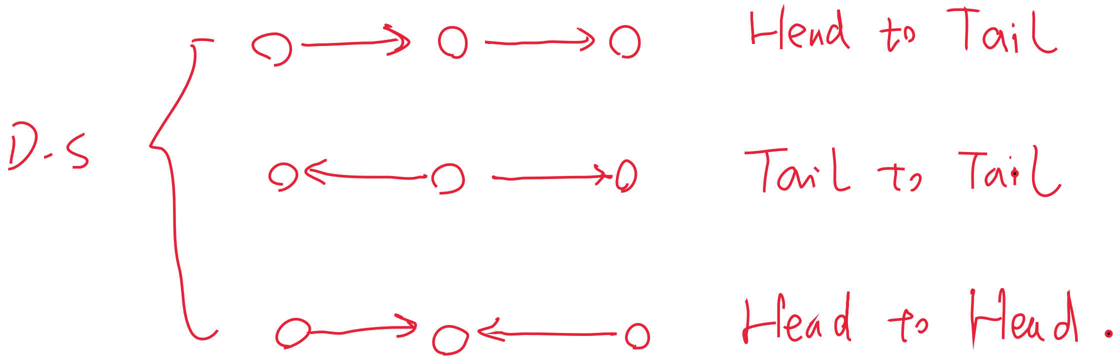
$X_1 \perp\!\!\!\perp X_4 \mid \{X_2, X_3\}$

$X_1 \not\perp\!\!\!\perp X_4 \mid X_3$



D-separation

Quiz.



EM Algorithm - Precisely

EM is a general procedure for learning from partly observed data

Given observed variables X, unobserved Z (X={F,A,H,N}, Z={S}) ✓

Define $Q(\theta'|\theta) = E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

next step current current M step new

Iterate until convergence:

- E Step: Use X and current θ to calculate P(Z|X,θ) $= \frac{P(X,Z|\theta)}{\sum_z P(X,Z=z|\theta)}$
- M Step: Replace current θ by

$$\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$$

$$Q(\theta'|\theta) = \sum_z P(Z=z|X,\theta) \log P(X,Z=z|\theta')$$

$$\theta' = \{\theta'_f, \dots, \theta'_n\}$$

$$\frac{\partial Q}{\partial \theta_f} = 0 \Rightarrow \theta_f =$$

$$\vdots$$

$$\frac{\partial Q}{\partial \theta_s} = 0 \Rightarrow \theta_s =$$

Guaranteed to find local maximum.
Each iteration increases $E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

Q

Sample Complexity for Supervised Learning

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\dots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

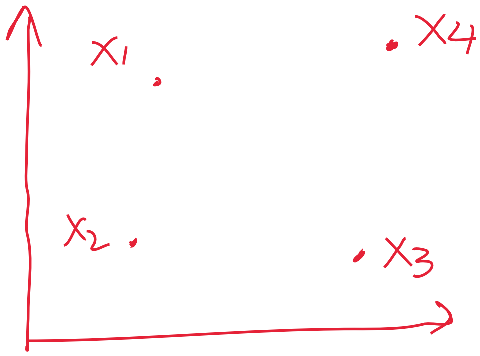
Theorem

$$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right]$$ Quiz

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Contrapositive: if the target is in H, and we have an algo that can find consistent fns, then we only need this many examples to get generalization error $\leq \epsilon$ with prob. $\geq 1-\delta$

Today's Quiz



$S = \{x_1, x_2, x_3, x_4\}$

$\boxed{H: \text{ quadratic separator}}$

① $H(S) \overset{?}{=}$

② $H[m] \overset{?}{=}$

③ $\mathrm{VC\,dim}(H) \geq 4$?

Analyzing Training Error: Proof Math

Step 1: unwrapping recurrence: $D_{T+1}(i) = \frac{1}{m}\left(\frac{\exp(-y_i f(x_i))}{\prod_t Z_t}\right)$ where $f(x_i) = \sum_t \alpha_t h_t(x_i)$.

Step 2: $\mathrm{err}_S(H_{final}) \leq \prod_t Z_t$.

Step 3: ${}_t Z_t = \prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} = {}_t\sqrt{1-4\gamma_t^2} \leq e^{-2\sum_t \gamma_t^2}$

Note: recall $Z_t = (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t} = 2\sqrt{\epsilon_t(1-\epsilon_t)}$

α_t minimizer of $\alpha \rightarrow (1-\epsilon_t)e^{-\alpha} + \epsilon_t e^{\alpha}$ **Quiz**

$err_D(g) \le err_S(g) + \tilde{O}\left(\sqrt{\frac{dT}{m}}\right)$

$err_D(g) = Pr_D(g(x) \ne y)$

$= Pr_D(H_f(x) \ne y)$

$err_S(g) = Pr_S(g(x) \ne y) \Rightarrow \underline{Pr_S(yf(x) \le 0)} \le Pr_S(yf(x) \le \theta)$

$(\theta > 0)$

Boosting and Margins

Theorem: VCdim$(H) = d$, then with prob. $\ge 1 - \delta$, $\forall f \in co(H)$, $\forall \underline{\theta > 0}$,

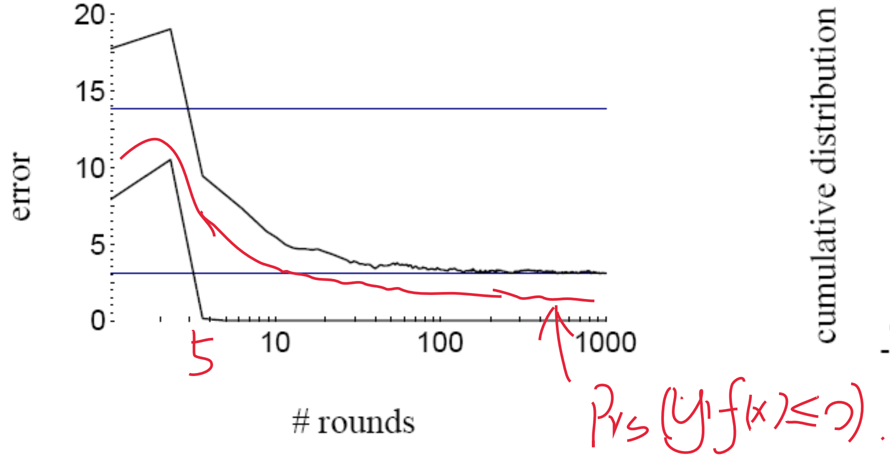
$H_f(x) \ne y$

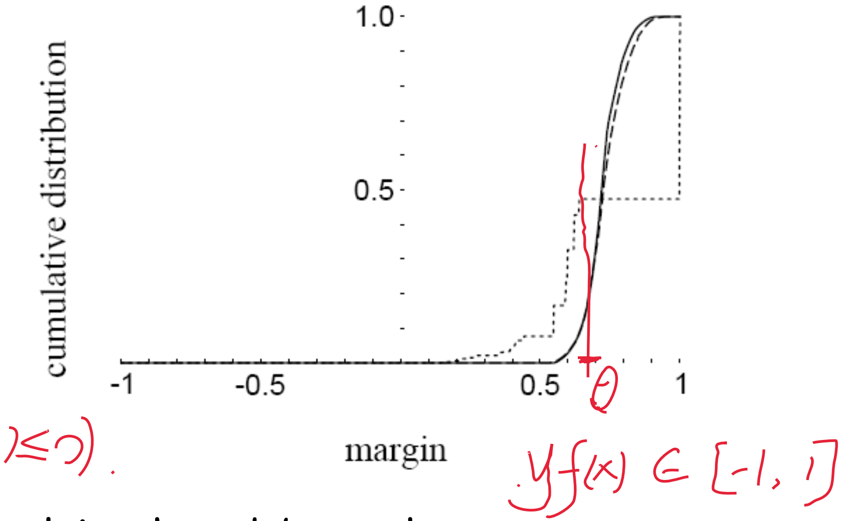
$$\Pr_D[\underline{yf(x) \le 0}] \le \boxed{\Pr_S[\underline{yf(x)} \le \underline{\theta}]} + \underline{O\left(\frac{1}{\sqrt{m}}\sqrt{\frac{d \ln^2 \frac{m}{d}}{\theta^2} + \ln\frac{1}{\delta}}\right)} = \tilde{O}\left(\sqrt{\frac{d}{m\theta^2}}\right)$$

threshold

Note: bound does **not** depend on T (the # of rounds of boosting), depends only on the complex. of the weak hyp space and the margin!

d θ

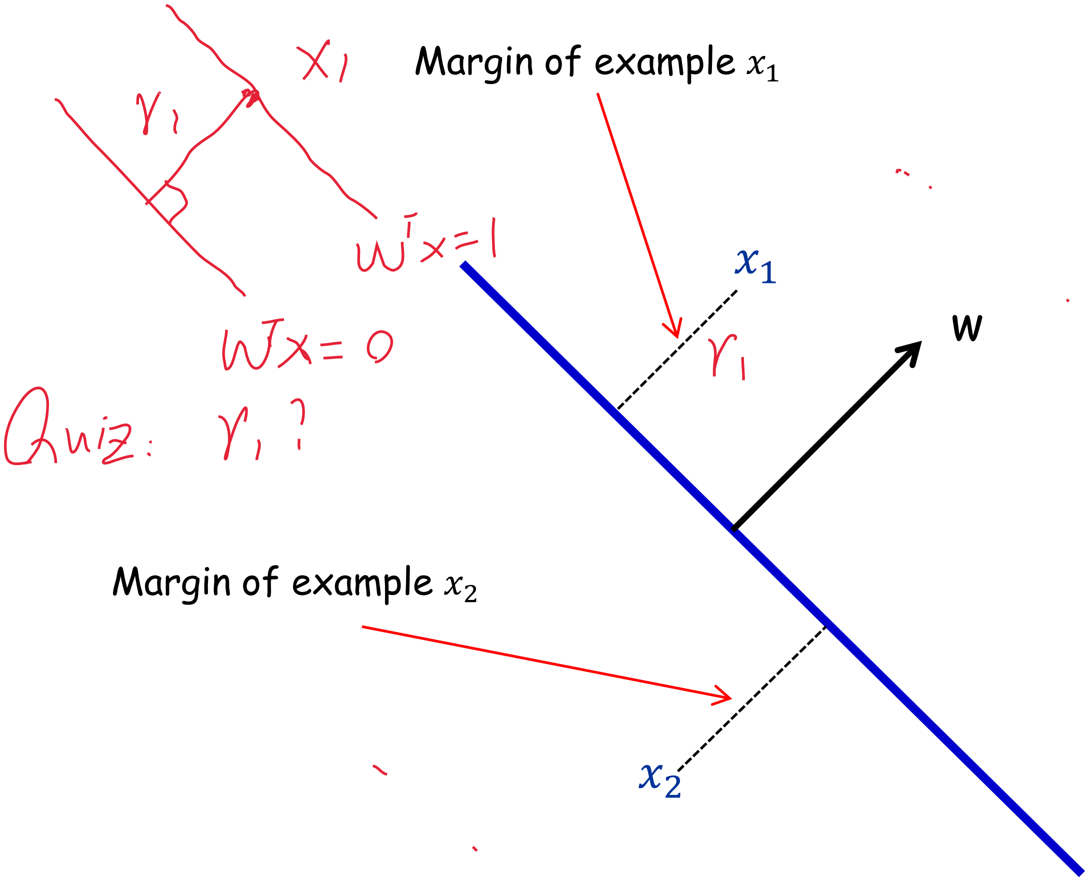




Quiz: according to this slide, explain why adaboost keeps decreasing testing error, even if training error equals to zero.

Geometric Margin

Definition: The margin of example x w.r.t. a linear sep. w is the distance from x to the plane $w \cdot x = 0$.



Margin of example x_1

Margin of example x_2

Quiz: γ_1 ?

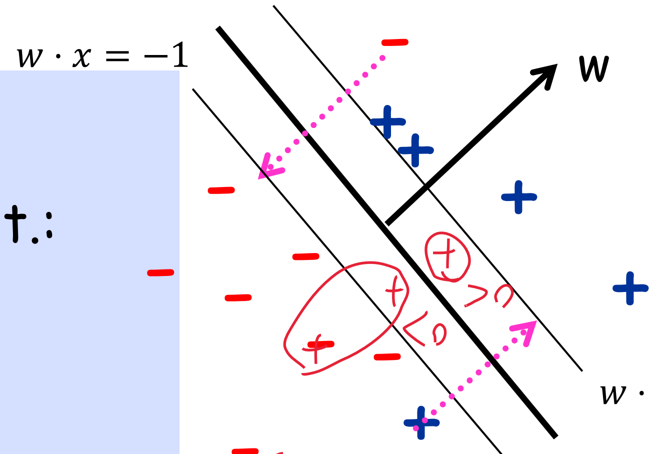
If $||w|| = 1$, margin of **x** w.r.t. **w** is $|x \cdot w|$.

$$\gamma_i = \frac{y_i w^T x_i}{||w||} \geq 0$$

Support Vector Machines (SVMs)

Question: what if data isn't perfectly linearly separable?

Replace "# mistakes" with upper bound called "hinge loss"



Input: S={(x_1, y_1), …,(x_m, y_m)};

Find $\text{argmin}_{w,\xi_1,\ldots,\xi_m} ||w||^2 + C \sum_i \xi_i$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1 - \xi_i$

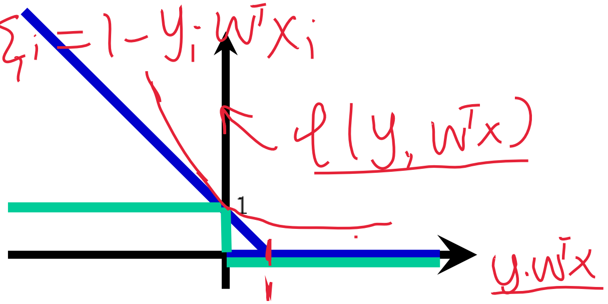
  $\xi_i \geq 0$

① $y_i w^T x_i \geq 1$, $\xi_i = 0$

$e^{(-y w^T x)}$ $f(x)$

ξ_i are "slack variables"

② $y_i w^T x_i < 1$, $\xi_i = 1 - y_i w^T x_i$



C controls the relative weighting between the twin goals of making the $||w||^2$ small (margin is large) and ensuring that most examples have functional margin ≥ 1.

Quiz

$l(w, x, y) = \max(0, 1 - y\, w \cdot x)$

hinge loss : $\xi_i = \max(0, 1 - y w^T x)$

$= (1 - y w^T x)_+$

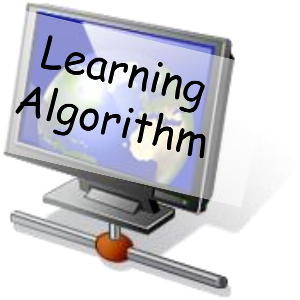
Modern ML: New Learning Approaches

Modern applications: massive amounts of raw data.

Techniques that best utilize data, minimizing need for expert/human intervention.

Paradigms where there has been great progress.

- Semi-supervised Learning, (Inter)active Learning.



Quiz

An Easy Case for k-means: k=1

Input: A set of n datapoints $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ in R^d

Output: $c \in \mathrm{R}^\mathrm{d}$ to minimize $\sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\left|\mathbf{x}^\mathrm{i} - \mathbf{c}\right|\right|^2$

Solution: The optimal choice is $\boldsymbol{\mu} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}^\mathrm{i}$

Idea: bias/variance like decomposition

$$\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\left|\mathbf{x}^\mathrm{i} - \mathbf{c}\right|\right|^2 = \left|\left|\boldsymbol{\mu} - \mathbf{c}\right|\right|^2 + \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\left|\mathbf{x}^\mathrm{i} - \boldsymbol{\mu}\right|\right|^2$$

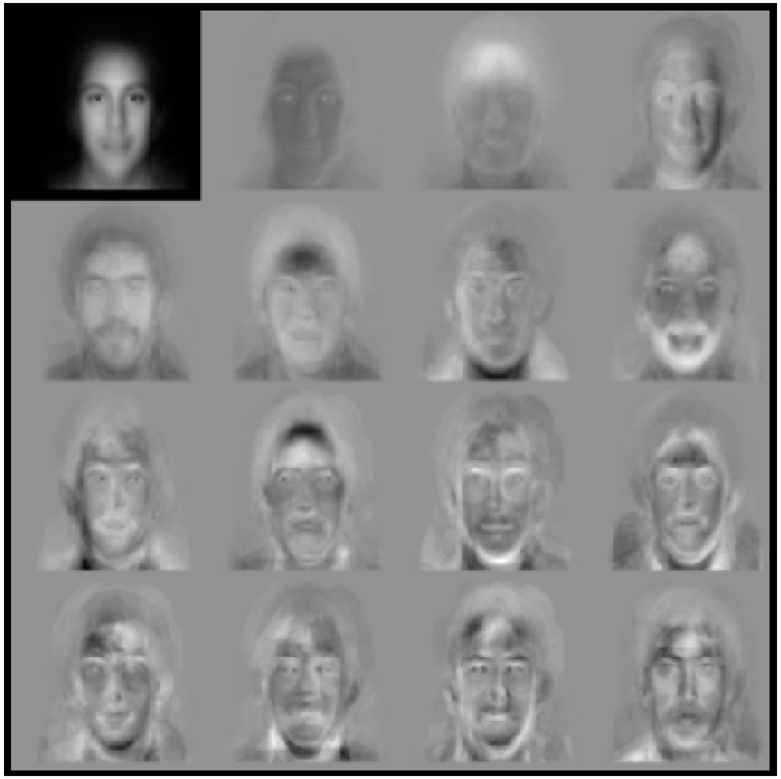
Quiz

Avg k-means cost wrt c

Avg k-means cost wrt μ

So, the optimal choice for $\mathbf{c}$ is $\boldsymbol{\mu}$.

Example: faces



Eigenfaces from 7562 images:

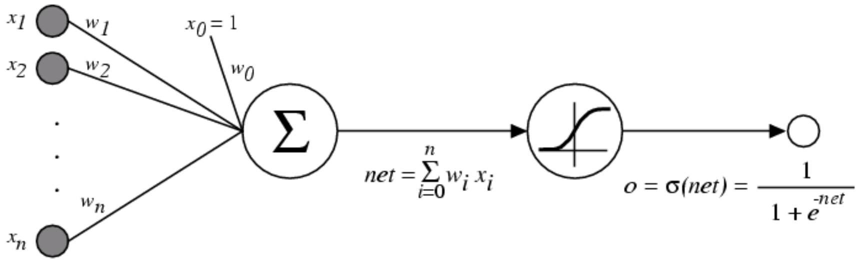
top left image is linear combination of rest.

Sirovich & Kirby (1987)
Turk & Pentland (1991)

Can represent a face image using just 15 numbers!

Quiz

Error Gradient for a Sigmoid Function


x_1
w_1
x_2
w_2
$x_0 = 1$
w_0
x_n
w_n
Σ
$net = \sum_{i=0}^{n} w_i x_i$
$o = \sigma(net) = \frac{1}{1 + e^{-net}}$

x_d = input

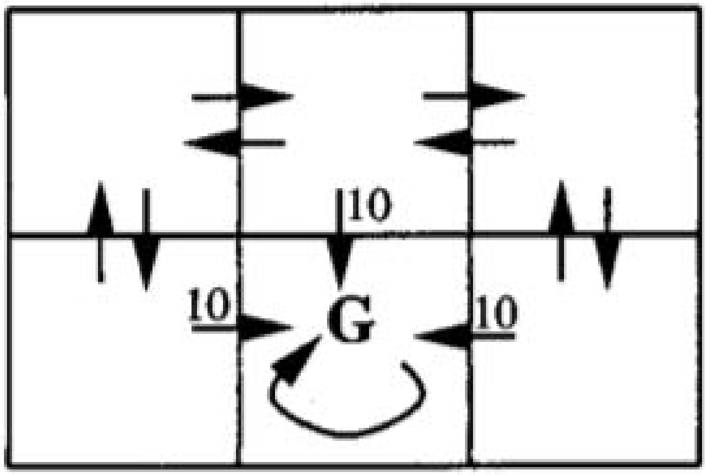
t_d = target output

o_d = observed unit output

w_i = weight i

Quiz

Quiz



\gamma=0.9

(1) Give an optimal policy for the above problem;
(2) Calculate the V*(s) values;
(3) Calculate the Q(s,a) values.

Positive semidefinite cone

notation:

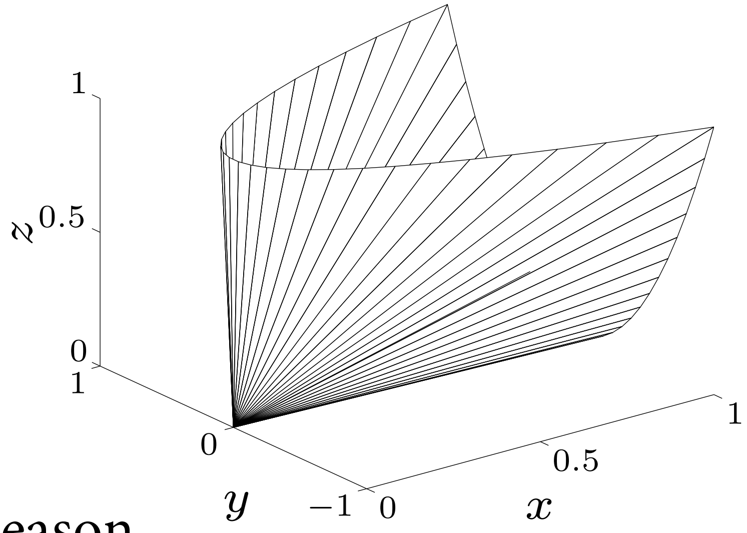
- $\mathbf{S}^n$ is set of symmetric $n \times n$ matrices
- $\mathbf{S}^n_+ = \{X \in \mathbf{S}^n \mid X \succeq 0\}$: positive semidefinite $n \times n$ matrices

$$X \in \mathbf{S}^n_+ \quad \Longleftrightarrow \quad z^T X z \geq 0 \text{ for all } z$$

  $\mathbf{S}^n_+$ is a convex cone

- $\mathbf{S}^n_{++} = \{X \in \mathbf{S}^n \mid X \succ 0\}$: positive definite $n \times n$ matrices

example: $\begin{bmatrix} x & y \\ y & z \end{bmatrix} \in \mathbf{S}^2_+$



Quiz:

Is S^n_{++} a convex cone? Show the reason.

Affine function

suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ is affine ($f(x) = Ax + b$ with $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$)

- the image of a convex set under f is convex

$$S \subseteq \mathbf{R}^n \text{ convex} \quad \Longrightarrow \quad f(S) = \{f(x) \mid x \in S\} \text{ convex}$$

- the inverse image $f^{-1}(C)$ of a convex set under f is convex

Quiz $$C \subseteq \mathbf{R}^m \text{ convex} \quad \Longrightarrow \quad f^{-1}(C) = \{x \in \mathbf{R}^n \mid f(x) \in C\} \text{ convex}$$

Pointwise supremum

if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is convex **Quiz**

examples

- support function of a set C: $S_C(x) = \sup_{y \in C} y^T x$ is convex
- distance to farthest point in a set C:

$$f(x) = \sup_{y \in C} \|x - y\|$$

- maximum eigenvalue of symmetric matrix: for $X \in \mathbf{S}^n$,

$$\lambda_{\max}(X) = \sup_{\|y\|_2 = 1} y^T X y$$

Log-concave and log-convex functions

a positive function f is log-concave if $\log f$ is concave:

$$f(\theta x + (1-\theta)y) \geq f(x)^{\theta} f(y)^{1-\theta} \quad \text{for } 0 \leq \theta \leq 1$$

f is log-convex if $\log f$ is convex

- powers: x^a on $\mathbf{R}_{++}$ is log-convex for $a \leq 0$, log-concave for $a \geq 0$
- many common probability densities are log-concave, *e.g.*, normal:

Quiz

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\bar{x})^T \Sigma^{-1}(x-\bar{x})}$$

- cumulative Gaussian distribution function Φ is log-concave

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

LP and SOCP as SDP

LP and equivalent SDP

LP: $\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$ SDP: $\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathbf{diag}(Ax - b) \preceq 0 \end{array}$

(note different interpretation of generalized inequality $\preceq$)

SOCP and equivalent SDP

SOCP: $\begin{array}{ll} \text{minimize} & f^T x \\ \text{subject to} & \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m \end{array}$

SDP: $\begin{array}{ll} \text{minimize} & f^T x \\ \text{subject to} & \left[\begin{array}{cc} (c_i^T x + d_i)I & A_i x + b_i \\ (A_i x + b_i)^T & c_i^T x + d_i \end{array}\right] \succeq 0, \quad i = 1, \ldots, m \end{array}$

Quiz: how to represent QP as SDP?

Lagrange dual and conjugate function

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \preceq b, \quad Cx = d \end{array}$$

dual function

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathbf{dom}\, f_0} \left(f_0(x) + (A^T\lambda + C^T\nu)^T x - b^T\lambda - d^T\nu \right) \\ &= -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu \end{aligned}$$

- recall definition of conjugate $f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$
- simplifies derivation of dual if conjugate of f_0 is known

example: entropy maximization

$$f_0(x) = \sum_{i=1}^n x_i \log x_i, \qquad f_0^*(y) = \sum_{i=1}^n e^{y_i - 1}$$

Quiz: derive the dual function of entropy maximization problem.