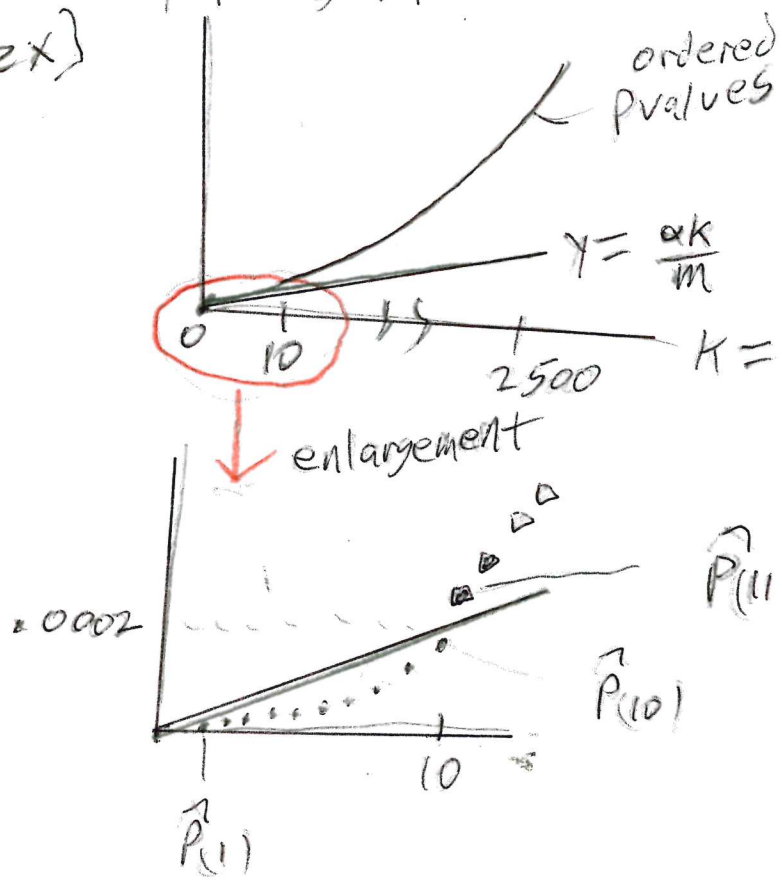
ex] Y = pvalues = P-values



$\frac{.05}{2500} = .00002$ Bonferroni Cutoff

$\hat{k} = 10$

$10\,\frac{(.05)}{2500} = .0002$

ex] $n = 102$ men: 52 prostate cancer patients, 50 normal controls.

Each man's gene expression levels were measured on a panel of 6033 genes: x_{ij} = activity of ith gene for jth man.

2 sample t statistic (pooled) $t_i = \frac{\bar{x}_{2i} - \bar{x}_{1i}}{\hat{\sigma}_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ was computed for gene i

2 = cancer, 1 = normal.

Then get $\hat{P}_1, \ldots, \hat{P}_m$ = pvalues for $i = 1, \ldots, m = 6033$.

Multivariate Linear Regression

Ch 10

1] For multivariate linear regression (mreg), there are $m \geq 2$ response variables $Y_1, \ldots, Y_m$. At least one predictor variable takes on many values.

ex] systolic and diastolic blood pressure

2] The mreg model is

$$\underset{m\times 1}{\underset{\sim}{y_i}} = \underset{m\times p}{B^T} \underset{p\times 1}{\underset{\sim}{x_i}} + \underset{m\times 1}{\underset{\sim}{\varepsilon_i}} \quad \text{for } i = 1, \ldots, n$$

with $m \geq 2$ response variables $Y_1, \ldots, Y_m$ and p predictor variables $x_1 \equiv 1, \underbrace{x_2, \ldots, x_p}_{\text{nontrivial predictors}}$. The

ith case $= (\underline{x}_i^T, \underline{y}_i^T)^T$ where

$x_{i1} = 1$ is often omitted.

data matrix $\begin{pmatrix} \underline{x}_1^T, \underline{y}_1^T \\ \vdots \\ \underline{x}_n^T, \underline{y}_n^T \end{pmatrix}$ but column of 1s is usually omitted.

In matrix form, $\underset{n\times m}{\mathbb{Z}_1} = \underset{n\times p}{\mathbb{X}}\,\underset{p\times m}{B} + \underset{n\times m}{E}$ where

$$\mathbb{Z}_1 = (\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_m) = \begin{pmatrix} \underline{z}_1^T \\ \vdots \\ \underline{z}_n^T \end{pmatrix},$$

$$B = [\underline{\beta}_1\ \underline{\beta}_2\ \cdots\ \underline{\beta}_m],\quad E = [\underline{e}_1, \ldots, \underline{e}_m] = \begin{pmatrix} \underline{\varepsilon}_1^T \\ \vdots \\ \underline{\varepsilon}_n^T \end{pmatrix}.$$

$E[\underline{\varepsilon}_k] = \underline{0}$, $\mathrm{cov}(\underline{\varepsilon}_k) = \Sigma_{\underset{\sim}{\varepsilon}} = (\sigma_{ij})$, $k = 1, \ldots, n$,

$E(\underline{e}_i) = \underline{0}$, $\mathrm{cov}(\underline{e}_i, \underline{e}_j) = \sigma_{ij} I_n$ for $i, j = 1, \ldots, m$.

3] Use $\hat{\underline{\varepsilon}}$, $\hat{\underset{\sim}{\Sigma}}$, and $\hat{\varepsilon}_{ij}$ for residuals.

4] * Each response variable Y_j in the mreg model follows a univariate multiple linear regression (MLR) model $\underline{Y}_j = \mathbb{X}\underline{\beta}_j + \underline{e}_j$, $j = 1, \ldots, m$.

$Y_{ij} = \underline{x}_i^T \underline{\beta}_j + e_{ij} = \beta_{1j} + x_{i2}\beta_{2j} + \cdots + x_{ip}\beta_{pj} + e_{ij}$

with $x_{i1} \equiv 1$. $\mathbb{X}$ does not depend on j. So the same predictors are used for all m response variables.

5] $m = 1 \rightarrow$ MLR. So for mreg, $m \geq 2$.

6] The OLS estimator is

$\hat{\mathbf{B}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Z}_1$ which gives

$\hat{\underline{\beta}}_j = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underline{Y}_j$ $j = 1, \ldots, m$.

So $\hat{\mathbf{B}}$ can be computed by

computing the OLS MLR model for regressing $\underset{\sim}{Y}_j$ on $\underset{\sim}{X}$, $j=1,\ldots,m$.

7} competing estimators include

A} Regress $\underset{\sim}{Y}_j$ on $\underset{\sim}{X}$, $j=1,\ldots,m$

to get $\hat{\underset{\sim}{B}}_{jE}$, $j=1,\ldots,m$ and

take $\hat{B} = [\hat{\underset{\sim}{B}}_{1E}, \ldots, \hat{\underset{\sim}{B}}_{mE}]$

B} Regress $\underset{\sim}{Y}_j$ on $\underset{\sim}{X}$, $j=1,\ldots,m$

and get linear combinations

$\hat{\underset{\sim}{n}}_{11}^T \underset{\sim}{X}, \ldots, \hat{\underset{\sim}{n}}_{1K_1}^T \underset{\sim}{X}$

$\hat{\underset{\sim}{n}}_{21}^T \underset{\sim}{X}, \ldots, \hat{\underset{\sim}{n}}_{2K_2}^T \underset{\sim}{X}$

$\vdots$

$\hat{\underset{\sim}{n}}_{m1}^T \underset{\sim}{X}, \ldots, \hat{\underset{\sim}{n}}_{mK_m}^T \underset{\sim}{X}$

$= W_1, \ldots, W_K$ with $K = \sum_{i=1}^{m} K_i$.

Then use $\underset{n\times m}{\underset{\sim}{Y}} = \underset{n\times(K+1)}{\underset{\sim}{W}} \overset{(K+1)\times m}{\underset{\sim}{B}} + \underset{n\times m}{\underset{\sim}{E}}$ and

$\underset{\sim}{W} = [\underset{\sim}{1}\ \underset{\sim}{W_1} \ldots \underset{\sim}{W_K}]$

use OLS mreg to get $\hat{B}$

if $n > \max\{1, (m+k)^2, mk+30, 10k\}$.

ex] RR a) could use

$$\hat{B}_R = (X^TX + \hat{\lambda} I_n)^{-1} X^T Z$$

$$= [X^TX + \hat{\lambda} I_n]^{-1} [X^T y_1 \cdots X^T y_m],$$

but then the same $\hat{\lambda}$ is used for all m regressions.

b) could get $\hat{\beta}_j = (X^TX + \hat{\lambda}_j I_n)^{-1} X^T y_j$

for $j = 1, \ldots, m$.

c) could get $\hat{\beta}_j$ as in a) or b),

form $w_j = \hat{\beta}_j^T x \qquad j = 1, \ldots, m$

and do OLS mreg of Z on

$$W = [1, w_1, \ldots, w_m]$$

8] mreg2 model

$$\underset{m\times 1}{\underline{y}_i} = \underset{m\times 1}{\underline{\alpha}} + \underset{m\times p}{B_S^T}\ \underset{p\times 1}{\underline{x}_i} + \underset{m\times 1}{\underline{\varepsilon}_i}$$

← nontrivial predictors

$$\underset{n\times m}{Z} = \begin{bmatrix} \underline{\alpha}^T \\ \vdots \\ \underline{\alpha}^T \end{bmatrix} + \underset{n\times p}{X_1}\ \underset{p\times m}{B_S} + \underset{n\times m}{E} \qquad (*)$$

$$= X \begin{pmatrix} \underline{\alpha}^T \\ B_S \end{pmatrix} + E = \underset{(p+1)\times m}{X\,B} + E$$

$$\underset{n\times(p+1)}{X} = (\underset{\sim}{1}\ X_1).$$

The ith row of (*) is $\underline{y}_i^T = \underline{\alpha}^T + \underline{x}_i^T B_S + \underline{\varepsilon}_i^T$.

9] OLS estimator $\hat{B} = (X^T X)^{-1} X^T Z$.

2nd way to compute OLS estimator

$$\hat{\underline{\alpha}} = \bar{\underline{y}} - \hat{B}_S^T \bar{\underline{x}}, \qquad \hat{B}_S = \hat{\Sigma}_{\underline{x}}^{-1}\ \hat{\Sigma}_{\underline{x}\underline{y}}.$$

10] If $\underline{v}_i = \begin{pmatrix} \underline{y}_i \\ \underline{x}_i \end{pmatrix}$ are iid, $E(\underline{v}) = \begin{pmatrix} E(\underline{y}) \\ E(\underline{x}) \end{pmatrix} = \begin{pmatrix} \underline{\mu}_y \\ \underline{\mu}_x \end{pmatrix}$,

$$\mathrm{cov}(\underline{v}) = \Sigma_{\underline{v}} = \begin{pmatrix} \Sigma_{\underline{y}} & \Sigma_{\underline{y}\underline{x}} \\ \Sigma_{\underline{x}\underline{y}} & \Sigma_{\underline{x}} \end{pmatrix}, \text{ then}$$

$$\underline{\alpha}_{OLS} = \underline{\alpha} = \underline{\mu}_y - B_S^T \underline{\mu}_x, \qquad B_S = \Sigma_{\underline{x}}^{-1}\ \Sigma_{\underline{x}\underline{y}} = B_{S,OLS}.$$

11} Suppose $\begin{pmatrix} \underline{y} \\ \underline{x} \end{pmatrix} \sim N_{m+p}\left[\begin{pmatrix} \underline{\mu}_y \\ \underline{\mu}_x \end{pmatrix}, \begin{pmatrix} \underline{\Sigma}_y & \underline{\Sigma}_{yx} \\ \underline{\Sigma}_{xy} & \underline{\Sigma}_x \end{pmatrix}\right]$

where $\underline{\Sigma}_x^{-1}$ exists.

Then $\underline{y}|\underline{x} \sim N_m(\underline{\mu}_{y|x}, \underline{\Sigma}_{y|x})$ where

$$\underline{\mu}_{y|x} = E(\underline{y}|\underline{x}) = \underline{\mu}_y + \underline{\Sigma}_{yx}\underline{\Sigma}_x^{-1}(\underline{x} - \underline{\mu}_x)$$

$$= \underline{\mu}_y - B_s^T \underline{\mu}_x + B_s^T \underline{x} \quad = \underline{\alpha} + B_s^T \underline{x},$$

and $\underline{\Sigma}_{y|x} = \mathrm{cov}(\underline{y}|\underline{x}) = \underline{\Sigma}_y - \underline{\Sigma}_{yx}\underline{\Sigma}_x^{-1}\underline{\Sigma}_{xy}$.

Hence $\underline{y}|\underline{x}$ follows the OLS mreg2 model.

12} Let A be a full rank $k \times p$ matrix with $1 \le k \le p$. Let $\underline{w} = A\underline{x}$.

If $\begin{pmatrix} \underline{y} \\ \underline{x} \end{pmatrix}$ is as in 11], then

$$\begin{pmatrix} \underline{y} \\ \underline{w} \end{pmatrix} \sim N_{m+k}\left[\begin{pmatrix} \underline{\mu}_y \\ \underline{\mu}_w \end{pmatrix}, \begin{pmatrix} \underline{\Sigma}_y & \underline{\Sigma}_{yw} \\ \underline{\Sigma}_{wy} & \underline{\Sigma}_w \end{pmatrix}\right],$$

and $y|\underline{w}$ follows an OLS mreg2 model as

in II} with $\underline{\mu}_W = \underline{\mu}_{Ax} = A\underline{\mu}_x$,

$\Sigma_W = \Sigma_{Ax} = A\Sigma_x A^T$; $\Sigma_{YW} = \Sigma_{Y,Ax} = \Sigma_{YX} A^T$,

$\Sigma_{WY} = \Sigma_{AxY} = A\Sigma_{xY}$.

multitude of models

$\hat{B} = [\text{diag } \hat{\Sigma}_x]^{-1} \hat{\Sigma}_{xY}$ does not fit data well

ex} OPLS type estimators

Let $w_1 = \hat{\Sigma}_{xY_1}^T \underline{x}$, ..., $w_m = \hat{\Sigma}_{xY_m}^T \underline{x}$.

a) Do SLR of Y_i on w_i to get

$\hat{\lambda}_i \hat{\Sigma}_{xY_i} = \hat{B}_{OPLS}(i)$.

can be done even if $n >> n$ and $p >> n$

$\hat{B}_S = [\hat{B}_{OPLS}(1), \ldots, \hat{B}_{OPLS}(m)]$.

b) Do mreg of Z_i on $(w_1, \ldots, w_m)^T = \underline{w}$

to get $\hat{B}_S = \hat{\Sigma}_{\underline{w}}^{-1} \hat{\Sigma}_{\underline{w}z}$

or $\hat{B} = (W^T W)^{-1} W^T Z_i$

want $n \geq \max(4m^2, m^2+30, 10m)$

$W = \begin{pmatrix} 1 & \underline{w}_1^T \\ \vdots & \\ 1 & \underline{w}_n^T \end{pmatrix}$

ex} PCR type estimators

Let $W_1 = \hat{e}_1^T \underline{x}$, ..., $W_p = \hat{e}_p^T \underline{x}$,

$\underline{W} = (W_1, \ldots, W_p)^T$.

regress $\underline{Y}$ on W_1, including a constant

then W_1, W_2

$\vdots$

$W_1, \ldots, W_K$ want $n \geq \max((m+1)^2, mK+30, 10K)$

Back to classification.

1} It has been observed that often

$\hat{\beta}_{OLS} \approx K \hat{\beta}_E$ for several estimators

such as $\hat{\beta}_E = \hat{\beta}_{LR}$ and $\hat{\beta}_E = \hat{\beta}_{PR}$.

2} Let $\pi_j = P(Y=j)$, $\underline{\mu}_j = E(\underline{X} | Y=j)$ and

Let $N_j = \#\, Y's = j$ for $j = 0, 1$.

Let $n = N_0 + N_1$. Then $\hat{\pi}_i = \frac{N_i}{n}$

and $\hat{\pi}_1 = 1 - \hat{\pi}_0$.

Let $\hat{\mu}_i = \bar{x}_i = \frac{1}{N_i} \sum_{j:y_j=i} x_j$ = sample mean of the x_j corresponding to $y_j = i$ for $i = 0, 1$.

Claim: $\hat{\Sigma}_{xy} = \hat{\pi}_1 \hat{\pi}_0 (\hat{\mu}_1 - \hat{\mu}_0)$.

Hence for iid cases,

$\beta = \Sigma_{xy} = \pi_1 \pi_0 (\mu_1 - \mu_0)$.

Proof] $\hat{\Sigma}_{xy} = \frac{1}{n} \sum_i x_i y_i - \bar{x}\bar{y}$

$= \frac{1}{n} \left[\sum_{j:y_j=1} x_j (1) + \sum_{j:y_j=0} x_j (0) \right] - \bar{x} \hat{\pi}_1$

$= \frac{1}{n} N_1 \hat{\mu}_1 - \frac{1}{n} \underbrace{(N_1 \hat{\mu}_1 + N_0 \hat{\mu}_0)}_{\sum_{i=1}^n x_i} \hat{\pi}_1$

$= \hat{\pi}_1 \hat{\mu}_1 - \hat{\pi}_1^2 \hat{\mu}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\mu}_0$

$= \hat{\pi}_1 \underbrace{(1 - \hat{\pi}_1)}_{\hat{\pi}_0} \hat{\mu}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\mu}_0 = \hat{\pi}_1 \hat{\pi}_0 (\hat{\mu}_1 - \hat{\mu}_0)$. □

3] If the (x_i, y_i) are iid,

compute $\hat{\beta}^T x_i = w_i$ for $i = 1, \ldots, n$.

plug w_i into a binary regression estimator

such as LR to get

$$\hat{\beta}_B = \hat{\lambda}\,\hat{\underline{n}}$$

(B: binary)

and do inference using $\sqrt{n}(\hat{\underline{n}} - \underline{n}) \xrightarrow{D} N_p(\underline{0}, \Sigma_{\underline{w}})$

as for MLR.

n_1 was N_1

4] suppose $\underline{x}_{11}, \ldots, \underline{x}_{1n_1}$ are iid

$\underline{x}_{01}, \ldots, \underline{x}_{0,n_0}$ are iid

and the 2 samples are ind.

Let $\underline{n} = \underline{\mu}_1 - \underline{\mu}_0$ and $\hat{\underline{n}} = \bar{\underline{x}}_1 - \bar{\underline{x}}_0$

Assume $n_i = \pi_i n$ and ($\frac{n_i}{n} \xrightarrow{P} \pi_i$ works)

$$\sqrt{n_i}\,(\bar{\underline{x}}_i - \underline{\mu}_i) \xrightarrow{D} N_p(\underline{0}, \Sigma_i)$$

$$\Sigma_i = \Sigma_{\underline{x}_i} = \mathrm{cov}(\underline{x}_i), \quad i = 0, 1.$$

So $\sqrt{n}\,(\bar{\underline{x}}_1 - \underline{\mu}_i) \xrightarrow{D} N_p(\underline{0}, \frac{\Sigma_i}{\pi_i})$ and

$$\sqrt{n}\begin{pmatrix} \bar{\underline{x}}_1 - \underline{\mu}_1 \\ \bar{\underline{x}}_0 - \underline{\mu}_0 \end{pmatrix} \xrightarrow{D} N_{2p}\left[\underline{0}, \begin{pmatrix} \frac{\Sigma_1}{\pi_1} & \underline{0} \\ \underline{0} & \frac{\Sigma_0}{\pi_0} \end{pmatrix}\right]$$

Thus $\sqrt{n}\left[\bar{x}_1 - \bar{x}_0 - (\mu_1 - \mu_0)\right] \xrightarrow{D} N_p(0, \Sigma_w)$ (*)

where $\Sigma_w = \frac{\Sigma_1}{\pi_1} + \frac{\Sigma_0}{\pi_2}$ and

$\hat{\Sigma}_w = \frac{n\hat{\Sigma}_1}{n_1} + \frac{n\hat{\Sigma}_0}{n_2}$.

Take $\hat{\eta}_B = \bar{x}_1 - \bar{x}_0$ and $\eta = \mu_1 - \mu_0$.

Then $\sqrt{n}(\hat{\eta}_B - \eta) \xrightarrow{D} N_p(0, \Sigma_w)$.

Let $w_i = \hat{\eta}^T x_i$ for $i = 1, \ldots, n$.

Plug w_i into a binary reg estimator such as LR to get $\beta_B = \hat{\lambda}\hat{\eta}_B$ and do inference using (*).

PhD topic

5)

$\bar{x}_1 - \bar{x}_0$ good for classification

6)

$\mu_1 = \mu_0$

$\bar{x}_1 - \bar{x}_0$ bad for classification

7] If the model depends on $\underline{x}$ only through $h(\underline{x})$, $h: \mathbb{R}^p \to \mathbb{R}$, written $Y \perp\!\!\!\perp \underline{x} \mid h(\underline{x})$
then $h(\underline{x})$ is a sufficient predictor (SP)

and $\hat{h}(\underline{x})$ is an estimated sufficient predictor

ESP.

ex] $ESP = \hat{\underline{\beta}}^T \underline{x}$

ex] $ESP = \hat{\alpha} + \hat{\underline{\beta}}^T \underline{x}$

8] Logistic regression is used a lot in biostatistics and epidemiology where the focus is statistical inference ($n \geq 10p$). Support vector machines (SVMs) are used in machine learning where the goal is classification accuracy.

9] When $p >> n$, there is often a hyperplane

that perfectly separates the two groups.

The launching point for SVMs was finding the "optimal" separating hyperplane.

10] For 2 groups, use $z \in \{-1, 1\}$.

Let $SP = \beta_0 + \underline{\beta}^T \underline{x}$.

$\hat{w}$ was α

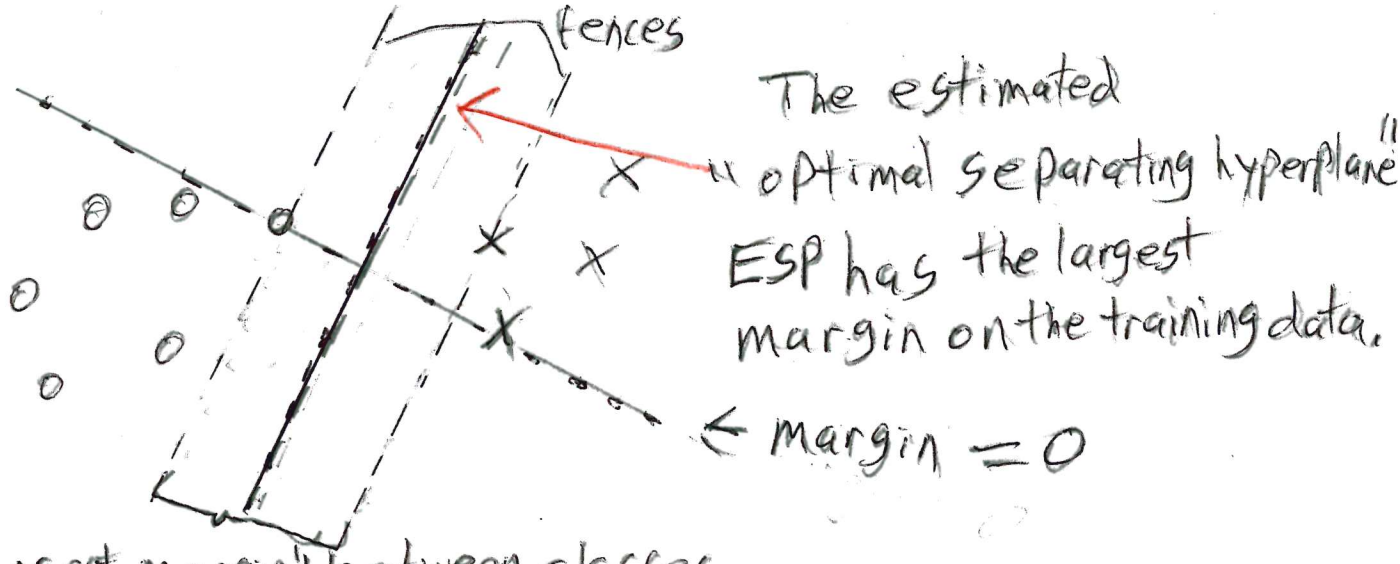
Classify $\underline{x}$ in group 1 if ESP > 0

-1 ESP < 0

(just like LR but code group 0 as -1).

So classifier $\hat{C}(\underline{x}) = \text{sign}(ESP)$.

11]



[2] The SVM split tries to make the 2 "halves" or partitions as homogeneous as possible.

[3] The hyperplanes parallel to the ESP that form the boundaries of the margin are called fences. The fences pass through at least 2 training data cases. These cases form the support set S of support vectors.

It turns out that

$\hat{\beta}_M = \sum_{i \in S} \hat{\alpha}_i \underset{\sim}{x}_i$ = optimal marginal classifier.

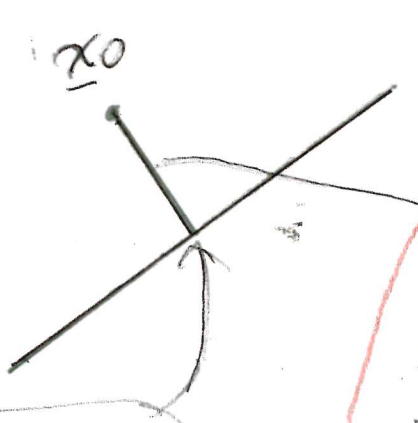
[4] Wide data = ultra high dimensional data has $p >> n$. If $n \leq p+1$ there is a separating hyperplane unless there are exact predictor ties across the class barrier (whatever that means).

15] Let $ESP = \hat{\beta}_0 + \underline{\hat{\beta}}^T \underline{x}$ and

$f(\underline{x}) = SP = \beta_0 + \underline{\beta}^T \underline{x} = SP.$

So $\hat{f}(\underline{x}) = ESP$, $z_i \in \{-1, 1\}$.

16]



projection of $\underline{x}_0$ onto hyperplane must be $\beta_0 + \underline{\beta}^T \underline{x}_0$

omit

signed distance of $\underline{x}_0$ from decision boundary $=$

$$\frac{\beta_0 + \underline{\beta}^T \underline{x}_0}{\|\underline{\beta}\|_2} = \frac{f(\underline{x}_0)}{\|\underline{\beta}\|_2}$$

17] Think of the hyperplane $\beta_0 + \underline{\beta}^T \underline{x}_i$

$= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ dividing $\mathbb{R}^p$ into 2 halves. A separating hyperplane has $SP > 0$ if $\underline{x} \in$ group 1

< 0 group -1.

So $z_i SP_i = z_i(\beta_0 + \underline{\beta}^T \underline{x}_i) > 0$ for $i = 1, \ldots, n$.

[18] Think of a binary classifier

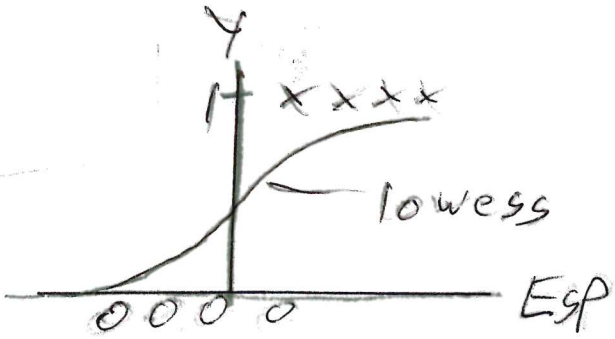
$y | \underline{x} \sim \text{bin}[m=1, g(\underline{x})]$ where

$g(\underline{x}) = g(SP) = P(Y=1 | \underline{x})$, where

$Z = -1$ is recorded as $y = 0$, but

$g(SP)$ is unknown.

response plot



Use lowess or a step function

[19] ← (see 11)

Let M be the margin. The optimal margin classifier $(\hat{\beta}_{0M}, \hat{\underline{\beta}}_M)$

maximizes M subject to

$\beta_0 \in \mathbb{R}, \underline{\beta} \in \mathbb{R}^p$

$z_i SP_i = z_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M$ (*)

$\forall\ i = 1, \ldots, n$.

This is called a hard margin classifier since no cases from either group can pass the fences of the classifier.

Equivalently, $\min_{\beta_0, \underline{\beta}} \|\underline{\beta}\|_2$ subject to (*).

20] A soft margin classifier allows cases from either group to pass the fences and thus be misclassified. This classifier solves $\min_{\beta_0, \underline{\beta}} \|\underline{\beta}\|_2$ subject to

$z_i (\beta_0 + \underline{x}_i^T \underline{\beta}) \geq 1 - \varepsilon_i$, $\varepsilon_i \geq 0$ for $i = 1, \ldots, n$

and $\sum_{i=1}^{n} \varepsilon_i \leq D$ some constant

ε_i are slack variables are used in linear programming.

21] This minimization is equivalent to

$$\min_{\beta_0, \underline{\beta}} \sum_{i=1}^{n} \underbrace{\left[1 - z_i (\beta_0 + \underline{x}_i^T \underline{\beta})\right]_+}_{\text{hinge loss}} + \lambda \|\underline{\beta}\|_2^2$$

$$\text{hinge loss} = \begin{cases} 0 & \text{if } \underline{x}_i \text{ is on the correct side} \\ \text{cost of } \underline{x}_i \text{ being on the wrong side of the margin} \end{cases}$$

where $[w]_+ = \begin{cases} w & w \geq 0 \\ 0 & w \leq 0 \end{cases}$.

This technique is similar to ridge reg,

22] A support vector machine that uses $\underline{x}_i$ minimizes the above criterion. For separable data, $(\hat{\beta}_{0,svm}, \underline{\beta}_{svm}) \xrightarrow{p} (\hat{\beta}_{0M}, \underline{\beta}_M)$ as $\lambda \to 0$.

23] It turns out that $\underline{\hat{\beta}}_{svm} = \sum_{i \in S} \hat{\gamma}_i \underline{x}_i$

and $\hat{\beta}_{0,svm} + \underline{x}^T \underline{\hat{\beta}}_{svm} = \hat{\beta}_{0,svm} + \sum_{i \in S} \hat{\gamma}_i \langle \underline{x}, \underline{x}_i \rangle$

where $\langle \underline{x}, \underline{x}_i \rangle = \underline{x}^T \underline{x}_i$. This quantity can be computed using the $n \times n$ Gram matrix $\mathbb{X}\mathbb{X}^T$ with $O(n^2 p)$ complexity or using $\mathbb{X}^T\mathbb{X}$ with $O(np^2)$ complexity.

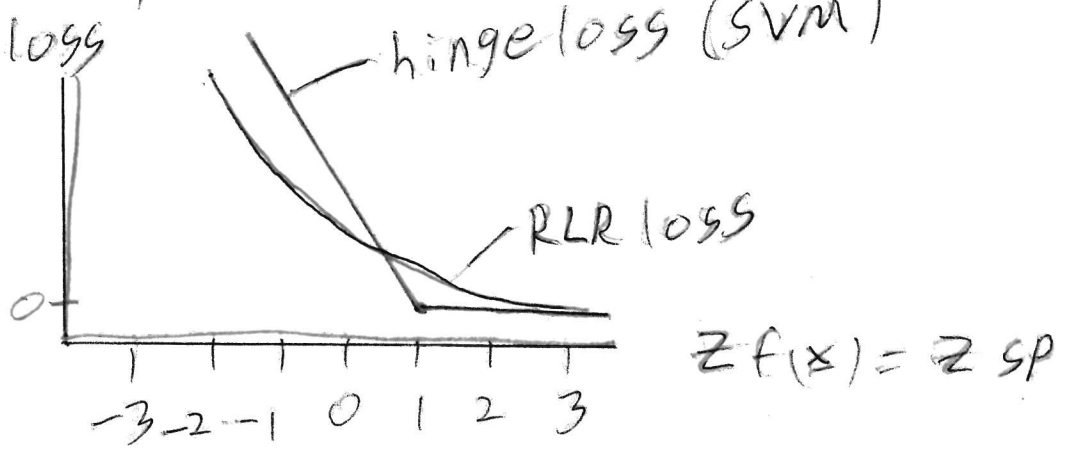
(RR used similar computations.)

24] A lasso-SVM solves

$$\min_{\beta_0, \underline{\beta}} \sum_{i=1}^{n} [1 - z_i(\beta_0 + \underline{x}_i^T \underline{\beta})]_+ + \underbrace{\lambda \|\underline{\beta}\|_1}_{\text{like lasso}}$$

and does variable selection.

25] "Ridged logistic regression" with $z_i \in \{-1, 1\}$ solves $\min_{\beta_0, \underline{\beta}} \sum_{i=1}^{n} \log[1 + e^{-z_i(\beta_0 + \underline{x}_i^T \underline{\beta})}] + \lambda \|\underline{\beta}\|_2^2$.



26] * Truth table = confusion matrix

predict \ truth	-1	1
-1	18	3
1	0	7

misclassified

correct

error rate $= 1 - \frac{18+7}{18+7+3}$

$= 1 - \frac{25}{28} = \frac{3}{28} = \boxed{0.1071}$

ex)

predict \ truth	a	b	c	d
a	10	0	0	12
b	0	100	11	0
c	0	5	50	0
d	6	0	0	30

diagonal: correctly classified
off diagonal: incorrectly classified

← correct

$$\text{error rate} = \frac{6+5+11+12}{6+5+11+12+10+100+50+30}$$

$$= \frac{\text{total} - \text{diagonal total}}{\text{total}} = \frac{224-190}{224}$$

$$= \frac{34}{224} = \boxed{0.1518}$$

See HW10

27) Sometimes 1 or a few observations shift the maximal margin hyperplane. The SVM classifier is a soft margin classifier and can do better.