$\boxed{\text{Recall}}$ Y data from model depending on β

(e.g. $Y = \beta X + \varepsilon$). Ingredients:

- prior distribution $\pi(\beta)$
- data likelihood $f(y|\beta)$
- posterior distribution $f(\beta|y) = \dfrac{f(y|\beta)\,\pi(\beta)}{f(y)}$

If $Y = X\beta + \varepsilon$ under A1, then $\underline{y} = (y_1, \dots, y_n)^T$ iid data here

$$f(\underline{y} \mid \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\underline{y} - \beta\underline{x})^T(\underline{y} - \beta\underline{x})\right)$$

depends on β

$$\beta = \#, \quad \underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Note: $(\underline{y} - \beta\underline{x})^T(\underline{y} - \beta\underline{x}) = \sum_{j=1}^{n} (y_j - \beta x_j)^2$

$\boxed{\text{Comment}}$ The posterior mode is the value of β that maximizes the posterior dist $f(\beta \mid \underline{y})$

maximize $f(\beta \mid \underline{y})$

$\iff$ maximize $\log f(\beta \mid \underline{y})$

$\Leftrightarrow$ minimize $-\log f(\beta | y)$

Thus the posterior mode minimizes

$$-\log f(\beta | y) = -\log\left[\frac{f(y|\beta)\pi(\beta)}{f(y)}\right] =$$

$$-\log f(y|\beta) - \log \pi(\beta) + \log f(y)$$ $\leftrightarrow$ does not depend on β

$$= \frac{1}{2\sigma^2}(y - \beta x)^T (y - \beta x) - \log \pi(\beta) \quad \left(+ \text{ constants that don't depend on } \beta \right)$$

Compare to ridge regression minimizes

$$(y - \beta x)^T (y - \beta x) + \lambda \beta^2$$

$\Rightarrow$ set $-\log \pi(\beta) = \lambda \beta^2$

$\Leftrightarrow$ $\underbrace{\pi(\beta) \propto e^{-\lambda\beta^2}}$ $\longrightarrow$ up to normalization constants, a normal pdf

$$c \cdot e^{-\frac{1}{2\left(\frac{1}{2\lambda}\right)}\left(\frac{\beta-0}{1}\right)^2}$$

The ridge regression estimator of β is the posterior mode in a Bayesian analysis using a normal, mean zero prior for β.

---

minimize

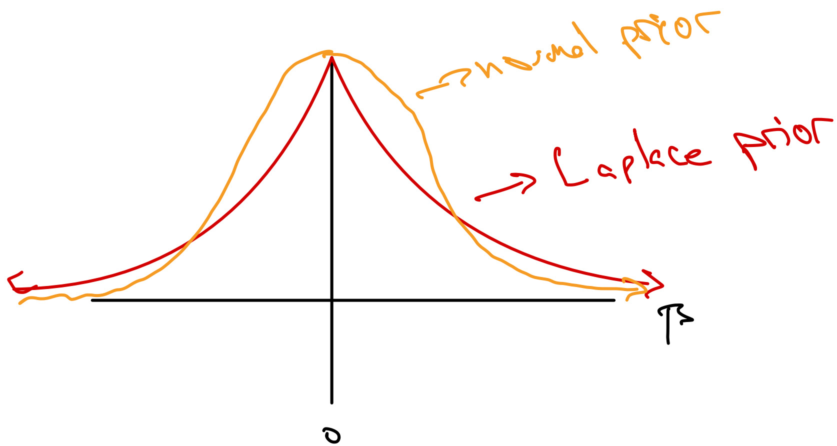
$$\frac{1}{2\sigma^2}\left[(\underline{y}-\beta\underline{x})^T(\underline{y}-\beta\underline{x})\right] - \log\pi(\beta) + \text{const}$$

Lasso:
minimize

$$(\underline{y}-\beta\underline{x})^T(\underline{y}-\beta\underline{x}) + \lambda|\beta|$$

$\Rightarrow \quad \pi(\beta) \propto e^{-\lambda|\beta|} \rightarrow$ double exponential or Laplace dist.

$\Rightarrow$ Lasso estimator is same as posterior mode under a mean Laplace prior.



"Elastic net"

$$(y - X\beta)^T (y - X\beta) + \overbrace{(1-\alpha)\|\beta\|_2^2}^{\text{ridge}} + \overbrace{\alpha\|\beta\|_1}^{\text{lasso}} \qquad \alpha \in [0,1]$$

7 Classification 2

$y_i \in \{-1, +1\}$ $\qquad \underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$

7.1 Maximum margin classifier

In p dimensions a <u>hyperplane</u> is flat subspace of dimension $(p-1)$

$\boxed{\text{DEF}}$ In p dim a <u>hyperplane</u> is the set of all

$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p$ that satisfy

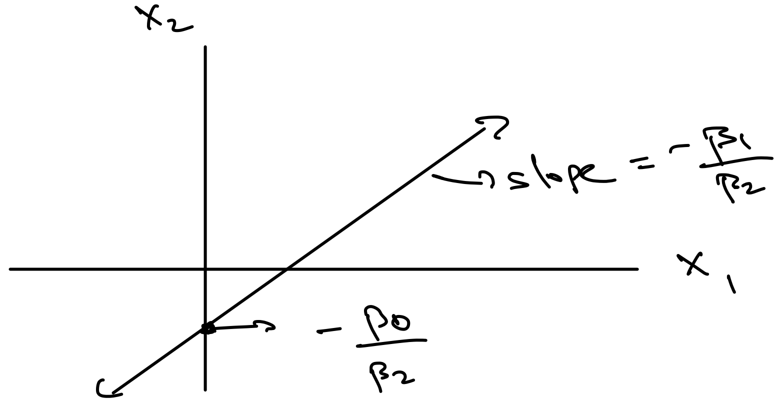
$$\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = 0 \qquad (\star)$$

for some parameters $\beta_0, \beta_1, \dots, \beta_p$

$\boxed{\text{Ex}}$ $p=2$ $\beta_0+\beta_1 x_1+\beta_2 x_2=0$

$$\rightarrow x_2=\left(-\frac{\beta_0}{\beta_2}\right)+\left(-\frac{\beta_1}{\beta_2}\right)x_1$$

$$= a + b x_1$$

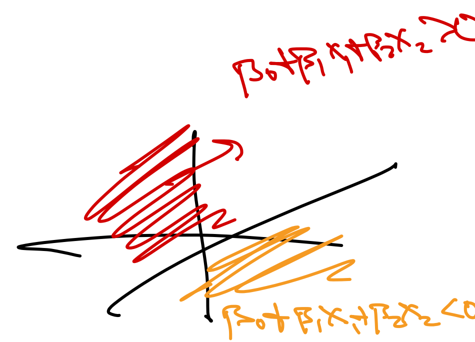


Note: If a vector $\underline{x}$ does not satisfy (*), then either

$$\beta_0+\beta_1 x_1+\dots+\beta_p x_p>0$$

$$\beta_0+\beta_1 x_1+\dots+\beta_p x_p<0$$

so a hyperplane splits space into halves



Setup Suppose we have training data $y_1, \ldots, y_n$ &

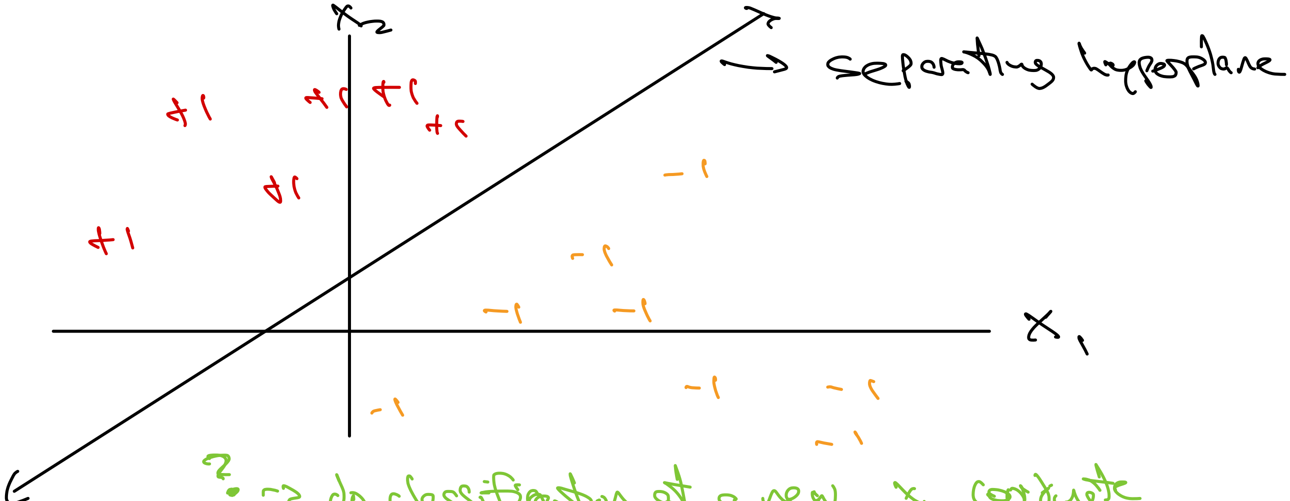
features $\underline{x}_1, \ldots, \underline{x}_n$ [does not include the 1 for intercept]

$$y_i \in \{-1, +1\} \qquad \underline{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \rightarrow \text{p-variate}$$

Goal Find a <u>separating hyperplane</u> such that

$$\begin{cases} \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0 & \text{if } y_i = +1 \\ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0 & \text{if } y_i = -1 \end{cases}$$

→ Collapse $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0$ for all i



?: → do classification at a new $\underline{x}$ coordinate

Suppose $\beta_0, \ldots, \beta_p$ are known (estimated), then at a new feature $\underline{x}_* = (x_{*1}, \ldots, x_{*p})^T$, our classifier is:

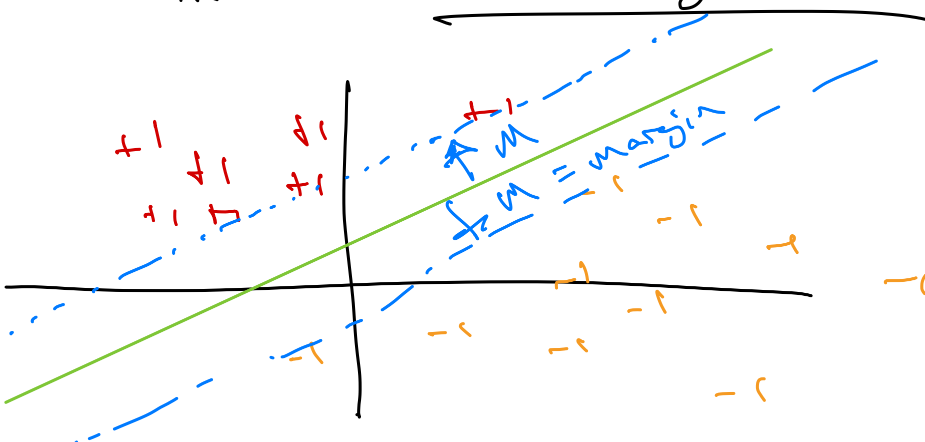
$$\hat{y} = \begin{cases} +1 & f(\underline{x}_*) > 0 \\ -1 & f(\underline{x}_*) < 0 \end{cases}$$

where $f(\underline{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

We can interpret the size/magnitude of f as confidence in our predictor.

Separable case

Suppose data are linearly separable. The maximum margin hyperplane is the plane that separates the classes but is as far away as possible from data. Yields the maximum margin classifier.



Note: The m.m. hy. is calculated by

- $\underset{\beta_0, \beta_1, \dots, \beta_p}{\text{maximize}} \quad M$

- subject to $\sum_{j=1}^{p} \beta_j^2 = 1$

  and $y_i(\beta_0 + x_i^T \beta) \geq M$ for all $i = 1, \dots, n$