2/12/15 Projections

<u>Linear regression</u>: minimizing sum of squares of errors in

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

i.e. find β s.t. $\|y - X\beta\|$ is minimized

solution by differentiating wrt β:

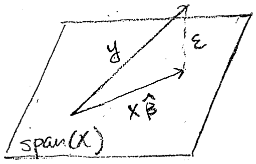
$\hat\beta = (X^TX)^{-1}X^Ty$ OLS estimator for β

X^TX invertible iff X has linearly indep. cols

fails when (1) more variables than samples

(2). perfect multicolinearity

<u>geometric interpretation of OLS</u>



$X\beta$ is a linear combination of cols of X

if $y \in \mathrm{span}(X)$, $\varepsilon = 0$

otherwise, ε is minimized when $X\hat\beta \perp \varepsilon$, i.e. $X\hat\beta$ is the orthogonal projection of y onto span(X)

<u>Projection matrix</u>: a square matrix $P: W \to W$ s.t. $P^2 = P$ (think geometrically)

in particular, $\forall v \in W$, $P(v - Pv) = 0$ $\quad W = \mathrm{Im}(P) \oplus \underbrace{\mathrm{Null}(P)}_{\text{"ker}(P)\text{"}}$ unique decomposition

<u>orthogonal projection</u>: $\mathrm{Im}(P)$ and $\mathrm{Null}(P)$ are orthogonal subspaces $\Leftrightarrow$ P is self-adjoint, $P = \overline{P^T}$

$P = P^T$

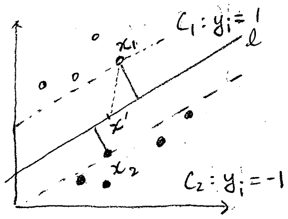
in particular, $Pv \perp v - Pv \quad \forall v \in W$

solving for the orthogonal projection $P: y \to \mathrm{span}(X)$

$Py = X\hat\beta$ for some $\hat\beta$ and $y - Py \perp \mathrm{span}(X) \Rightarrow X^T(y - X\hat\beta) = 0 = X^Ty - X^TX\hat\beta \Rightarrow \hat\beta = (X^TX)^{-1}X^Ty$

$\boxed{\text{PCA}}$

<u>SVM</u> find hyperplane that maximizes the separation between the 2 classes



training data

$\{(x_i, y_i) \mid x_i \in \mathbb{R}^k,$
$y_i \in \{1, -1\}\}$
$1 \le i \le n$

(1). <u>Hard margin</u>: hyperplane exists that separates the classes

- hyperplane ℓ represented by $w_1 x_{(i,1)} + \ldots + w_k x_{(i,k)} + b = 0$, or $w^Tx + b = 0$ $\begin{cases} > 0 & x_i \in C_1 \\ < 0 & x_i \in C_2 \end{cases}$
- let $x_1 \in C_1$, $x_2 \in C_2$ be the points closest to ℓ. WLOG, $w^Tx_1 + b = 1$ and $w^Tx_2 + b = -1$

$\Rightarrow y_i(w^Tx_i + b) \ge 1$

- objective is to maximize the margin around ℓ, i.e. $d(x_1, \ell) + d(x_2, \ell)$, where $d(x, \ell)$ is the projection of $x - x'$ onto the normal vector of ℓ, where x' is any pt on ℓ

to see that w is the normal vector: $w^T(x' - x'') = -b + b = 0$ for $x', x'' \in \ell$

- from before, $\underbrace{\mathrm{Proj}_w}_{\text{technically span}(w)}(v) = w\underbrace{(w^Tw)}_{s}^{-1}\underbrace{w^Tv}_{s} = \dfrac{v \cdot w}{w \cdot w}w$ "vector projection"

with $\|\mathrm{Proj}_w(v)\| = \dfrac{v \cdot w}{w \cdot w}\|w\| = \dfrac{v \cdot w}{\|w\|}$ "scalar projection" (can also be derived from $a \cdot b = |a||b|\cos\theta$)

$\Rightarrow d(x_1, \ell) = |\mathrm{Proj}_w(x_1 - x')| = w \cdot (x_1 - x') / \|w\| = 1/\|w\| = d(x_2, \ell)$

$\Rightarrow$ maximize $2/\|w\|$, or minimize $\|w\|^2$, subject to $y_i(w^Tx_i + b) \ge 1$

(2). <u>soft margin</u>: no linear separability (if $\xi_i > 1$, misclass.)

- introduce slack variables $\xi_i \ge 0$ that measure the degree to which x_i is on the wrong side of the margin
- new objective function is to maximize the margin while minimizing the classes overlap:

$$\min_{w, \xi_i} \left(\tfrac{1}{2}\right)\|w\|^2 + \left(\tfrac{C}{n}\right)\sum \xi_i \quad \text{subject to } y_i(w^Tx_i + b) \ge 1 - \xi_i,\ \xi_i \ge 0$$

$\xi_i \ge 1 - y_i(w^Tx_i + b)$ minimized when $\xi_i = 1 - y_i(w^Tx_i + b)$

$\Updownarrow$

$$\min_{w, b} \tfrac{1}{2}\|w\|^2 + \tfrac{C}{n}\sum\underbrace{(1 - y_i(w^Tx_i + b))}_{= \xi_i \ge 0,\ \text{so} = [1 - y_i(w^Tx_i + b)]_+} = \min_{w, b} \underbrace{\tfrac{1}{2}\|w\|^2}_{\ell^2\text{-reg. penalty}} + \underbrace{\tfrac{C}{n}\sum[1 - y_i(w^Tx_i + b)]_+}_{\text{hinge loss}}$$