Linear Regression Models.

<u>Assumptions:</u> (x_i, y_i) are IID's.

$$y_i = \underbrace{E(y_i \mid x_i)}_{f(x)} + \epsilon_i.$$

$E(\epsilon_i) = 0$

$$f(x) = \beta_0 + \beta_1 x.$$

$$M.S.E = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

$$\min_{\beta_1, \beta_0} J(\beta_1, \beta_0) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

$$\min_{u} J(u) = (Y - Au)^T (Y - Au) \quad = \text{least sq. Problem}$$

$$\nabla_u J(u) = A^T (Y - Au)$$

Normal equation $J(u) = (A^T A)^{-1} A^T Y.$

$$SSE = \sum_{i=1}^{k} (y_i - f(x_i))^2.$$

$$RMSE = \sqrt{\frac{1}{k} \times S.S.E.}$$

$$MAE = \frac{1}{k} \sum_{i=1}^{k} |y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{k} \left| \frac{MAE}{y_i} \right| \times 100.$$

$$NMSE = \frac{SSE}{(y_i - \bar{y})^2 = SS.T} = \frac{SSE}{SST}$$

$$R^2 = \sum_{i=1}^{k} (\hat{y}_i - \bar{\hat{y}}) / SS.T$$

Non-Linear Regression Models:

Quadratic fit: $\beta_0 + \beta_1 x + \beta_2 x^2$

If more than one feature

x_1, x_2

In degree 1, x_1, x_2.

2nd " , $x_1^2, x_2^2, x_1 \times x_2$

3rd " , $x_1^3, x_2^3, x_1^2 x_2, x_1 x_2^2, x_1 x_2$,

$x_1, x_2, x_1^2, x_2^2, 1$.

If we increase Model complexity chances of overfitting are there so regularization is needed.

$$\min_u (Y - Au)^T (Y - Au) + \lambda u^T u.$$ L_2-Regularization Ridge.

$$\nabla_u = -2A^T(Y - Au) + 2\lambda u = 0.$$

$$- A^T Y + A^T A u + \lambda u = 0$$

$$\Rightarrow - A^T Y + A^T A u + \lambda I u = 0$$

$$\Rightarrow (A^T A + \lambda I) u = A^T Y$$

$$\Rightarrow u = (A^T A + \lambda I)^{-1} A^T Y.$$

$$\min_{(w,b)} \frac{\lambda}{2} w^T w + \sum_{i=1}^{l} (y_i - (w^T x_i + b))^2$$

Formulae: $\frac{(n+d)!}{n!\,d!}$ $\left\{ \begin{array}{l} n = \text{features} \\ d = \text{dimension} \end{array} \right\}$

Quadratic fn in R^n.

$$\begin{bmatrix} w_m x_n^2 + \cdots \; \beta_n \\ w_{m-1} x_{n-1}^2 \end{bmatrix}$$

$$w_m \phi_m(x) + w_{m-1} \phi_{m-1}(x) + \cdots w_1 \phi_1(x) + b\phi(x)$$

$$f(x) = w^T \phi(x) + b$$

$$w = \begin{bmatrix} w_m \\ w_{m-1} \\ \vdots \\ w_1 \\ b \end{bmatrix}, \quad \phi(x) = \begin{bmatrix} \phi_m(x) \\ \phi_{m-1}(x) \\ \vdots \\ \phi_1(x) \\ 1 \end{bmatrix}$$

$$\min_{(w,b)} \frac{\lambda}{2} w^T w + \sum_{i=1}^{l} (y_i - w^T \phi(x) + b))^2$$

Basis functions:

R.B.F: $\phi_j(x) = \exp\left(-\frac{1}{2\sigma^2} \|x - u_j\|^2\right)$

$\phi_j(x) = e^{-\frac{1}{2\sigma^2}(\|x - u_j\|^2)}$

Maximum likelyhood estimate:

$$y_i = f(x_i) + \epsilon_i$$

Assumption: $(y|x_i)$ is from $N(E(y|x_i), \sigma^2)$

$N(w^T \phi(x_i), \sigma^2)$

$f(x_i) = w^T \phi(x_i) + \epsilon_i$ (In least squ.)

$y = w^T \phi(x_i) + \epsilon_i$

$\Rightarrow \epsilon_i = y - w^T \phi(x_i)$

$N(0, \sigma^2)$.

Selected sample $T = \{(x_i, y_i) : x_i \in R^n, y_i \in R\}$

Optimization problem.

$$\max p((y_1, y_2 \cdots y_l) | x_1, x_2 \cdots x_l)$$

$$= \max_{w} \prod_{i=1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2}}$$

$$\max_{w} \log\left(\prod_{i=1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2}} \right)$$

$$= \max_{w} \left(\sum_{i=1}^{l} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2}} \right) \right)$$

$$\max_{w} \sum_{i=1}^{l} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \sum_{i=1}^{l} \frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2}$$

$$\max_{w} - \sum_{i=1}^{l} (y_i - w^T\phi(x_i))^2$$

$$\min_{w} \sum_{i=1}^{l} (y_i - w^T\phi(x_i))^2$$

Quiz 1(1): $y = f(x) + \epsilon_i$.

$f(x) = w^T\phi(x_i) + b$.

$y = w^T\phi(x_i) + b + \epsilon_i \sim N(f(x), \sigma^2)$

$\Rightarrow \epsilon_i = y - w^T\phi(x_i) + b \sim N(0, \sigma^2)$

$$\max_{w} \prod_{i=1}^{l} \frac{1}{2\sigma} e^{-|y_i - (w^T\phi(x_i) + b)|}$$

$$\max_{w} \log\left(\prod_{i=1}^{l} \frac{1}{2\sigma} e^{-\frac{|y_i - w^T\phi(x_i) + b|}{\sigma}} \right)$$

$$= \max_{w} \sum_{i=1}^{l} \log\left(\frac{1}{2\sigma} \right) - \sum_{i=1}^{l} \frac{|y_i - (w^T\phi(x_i) + b)|}{\sigma}$$

$$\min_{w} |y_i - (w^T\phi(x_i) + b)|$$

Least Square Gradient Descent:

$$\min_{w} L(w) = \min_{w} \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i : w))^2.$$

$$\nabla_w L(w) = \cancel{2} \times \frac{1}{\cancel{2}} \sum_{i=1}^{n} (y_i - f(x_i : w))$$

$$= \sum_{i=1}^{n} (y_i - f(x_i : w)) \times \nabla_w f(x_i : w) = 0.$$

Algorithm: Step (1) Initialize w^0 to random pt.

Step 2. while $f(w^t)$ not converged do

$w^{(t+1)} \leftarrow w^t - \alpha_t \nabla f(w^t)$

Polynomial basis function:

$f(x) = w^T\phi(x)$, $w \in \mathbb{R}^{\frac{(n+d)!}{n!d!}}$

where $\phi(x) = \begin{bmatrix} \phi_m(x) \\ \vdots \\ \phi_1(x) \\ \phi_0(x) \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2^2 \\ \vdots \\ x_n \\ 1 \end{bmatrix}$

Gaussian: $\phi_j(x) = e^{-\frac{1}{2\sigma}\|x - \mu_j\|^2}$.

Sigmoidal = $\phi_j(x) = \sigma(x, w_j, b_j) = \frac{e^{w_j^T x + b_j}}{1 + e^{w_j^T x + b_j}}$

Kernel function & matrix

$\langle w \cdot x \rangle + b$

$K(x_i \cdot x_j) = \langle \phi(x_i) \cdot \phi(x_j) \rangle$

Kernel matrix (Gram Matrix):

① Positive definite matrix (All values > 0)

② Central structure ③ Fuse data and kernel

④ Information bottleneck.

$$\begin{bmatrix} K(1,1) & K(1,2) & K(1,3) & K(1,n) \\ K(2,1) & K(2,2) & K(2,3) & K(2,n) \\ K(m,1) & K(m,2) & K(m,3) & K(m,n) \end{bmatrix}$$

$K = (x_i^T x_j + 1)^M$ = Polynomial.

$K = (x^T x + 1)^M$

$K(x_i, x_j) = x_i^T x_j$ Linear.

R.B.F $= \exp\left(-\frac{1}{2\sigma^2} \|x_i - x_j\|^2\right)$

$\phi(a)^T \cdot \phi(b) = (a^T \cdot b)^2$

$=$

$$\sum_{i=1}^{l} K(x_i x)\alpha_i + b.$$

$$\frac{\lambda}{2} w^T w + \sum \left(y_i - (w^T x_i + b)\right)^2$$

Kernel Regression

Gradient descent least square kernel regression.

Step: 1) Initialize $x^0 = w^{start} \in R^M$ & $b \in R$.

Repeat.

$$w^{(j+1)} := w^j - \eta_k \left(\lambda w + \sum_{i=1}^{l} \frac{\partial s(w, b, x_k, y_k)}{\partial w}\right)$$

$$b^{(j+1)} := b^{(j)} - \eta_k \left(\sum_{i=1}^{l} \frac{\partial s(w, b, x_k, y_k)}{\partial b}\right)$$

for least square:

$$w^{(j+1)} := w^j - \eta_k \left(\lambda w + \sum_{i=1}^{l} -2\left(y_k - (w^T \phi(x_k) + b)\right)\phi\right)$$

$$b^{j+1} := b^j - \eta_k \left(\sum_{i=1}^{l} -2\left(y_k - (w^T \phi(x_k) + b)\right)\right)$$

Lasso Regression Model:

$$\min_w \frac{\lambda}{2} \|w\|_1 + \frac{1}{2} \sum_{i=1}^{l} \left(y_i - (w^T \phi(x_i))\right)^2.$$

Subgradient: A vector $g \in R^d$ is a subgradient of function $f: R^d \to R$ at a point $x \in R^d$ if $\forall z$ $f(z) \geq f(x) + g^T(z-x)$

$$g[i] = \begin{cases} = \text{sign}(w[i]), & \text{if } w[i] \neq 0 \\ \text{any element in } [-1,1], & \text{if } w[i] = 0 \end{cases}$$

Bias Variance:

Pattern Classification

Bayesian Decision Theory

$P(\omega_j|x)$ $\{\omega_j$ = A particular class$\}$

Discriminant function: $g_i(x)$ → to partition R^d space

Decision rule to assign x to ω_m

if $g_m(x) \geq g_i(x)$ $\forall i = 1,2 \ldots c$ & $i \neq m$

$g_k(x) = g_l(x)$ = Decision boundary

Linear discriminant fn:

$g_i(x) = w_i^T x + w_{0i}$.

Bayes Law:

$P(x|\omega_i)$ = likelihood → likelihood.

$P(\omega_j|x) = \dfrac{P(x|\omega_j) \cdot P(\omega_j)}{P(x)}$ → Prior; $P(x)$ → evidence.

↓ Posterior.

$\rightarrow P(\omega_j|x) = P(x|\omega_j) \cdot P(\omega_j)$

$P(x) = \sum_{j=1}^{2} P(x|\omega_j)P(\omega_j)$ ← Scale factor.

Choose ω_1 if $P(x|\omega_1) \cdot P(\omega_1) > P(x|\omega_2) \cdot P(\omega_2)$

else ω_2.

Probability of error: For a particular x.

$P(\text{error}|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{" " } \omega_1 \end{cases}$

To minimize.

Avg. prob of error = $P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}|x) P(x) \, dx$

or $\omega_{\text{best}} = \arg\max_j P(\omega_j|x) = \dfrac{P(x|\omega_j) P(\omega_j)}{P(x)}$ $j = 1,2,c$

loss = $\lambda(\alpha_i|\omega_j)$

if $g_i(x) > g_j(x)$ for all $j \neq i$.

$g_i(x) = P(\omega_i|x)$ or

$g_i(x) = P(x|\omega_i) P(\omega_i)$ or

$g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i)$

$g(x) = g_1(x) - g_2(x)$

Decide ω_1 if $g(x) > 0$;

otherwise decide ω_2.

$g(x) = p(\omega_1|x) - P(\omega_2|x)$

$g(x) = \ln p(x|\omega_1) p(\omega|\omega_2) + \ln p(\omega_1) k P(\omega_2)$

$g(x) = \ln p(x|\omega_i) + P(\omega_i)$

if $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$.

in this case we have.

$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$

$g(x) = g_1(x) - g_2(x)$

$\ln \dfrac{P(x|\omega_1)}{P(x|\omega_2)} + \ln \dfrac{P(\omega_1)}{P(\omega_2)}$

$P(x|\omega_i) = \dfrac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \times \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$

V

Case 1: $\Sigma_i = \sigma^2 I$ {only diagonal elements / Variance is same}

Hyper-spherical

$|\Sigma_i| = \sigma^{2d}$

$\Sigma_i^{-1} = \frac{1}{\sigma^2} \times I$ → Squared distance.

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln(P\omega_i)$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) + \ln(P(\omega_i))$$

$$= \frac{-1}{2\sigma^2}\Big[\underbrace{x^t x}_{\text{Indep. of } i} - 2\mu_i^t x + \mu_i^t \mu_i\Big] + \ln P(\omega_i)$$

$$= \frac{-1}{2\sigma^2}\left[-2\mu_i^t x + \mu_i^t \mu_i\right] + \ln P(\omega_i)$$

$= W_i^t x + W_{i0}$. — Linear Machine. For independent class.

$$W_i = \frac{1}{\sigma^2}\mu_i$$

$$w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i)$$

ω_i & ω_j → 2 classes.

$$g_i(x) - g_j(x) = 0$$

$$g_i(x) = w_i^t x + w_{i0}$$

$$g_j(x) = w_j^t x + w_{j0}$$

$$(w_i - w_j)^t x + w_{i0} - w_{j0} = 0.$$

$$\frac{1}{\sigma^2}(\mu_i - \mu_j)^T x - \frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(\omega_i) + \frac{\mu_j^t \mu_j}{2\sigma^2} - \ln P(\omega_j) = 0$$

$$= (\mu_i - \mu_j)^t x - \frac{1}{2}(\mu_i^t \mu_i - \mu_j^t \mu_j) + \sigma^2 \ln \frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$= (\mu_i - \mu_j)^t \left[x - \left\{\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} \times (\mu_i - \mu_j)\right\}\right] = 0$$

$$w^t(x - x_0) = 0$$

$$w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} \times (\mu_i - \mu_j)$$

Case 2: $\Sigma_i = \Sigma$ (There is co-variance but same)

Hyper Ellipse.

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$g_i(x) = w_i^t x + w_{i0}$$

$$w_i = \Sigma^{-1}\mu_i, \quad w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i)$$

$w^T(x - x_0) = 0$

$W = \Sigma^{-1}(\mu_i - \mu_j)$

$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_i)^T \Sigma^{-1}(\mu_i - \mu_j)} \cdot (\mu_i - \mu_j)$

Case 3: Σ_i = Arbitrary.

$g_i(x) = x^T A_i x + B_i^T x + C_{i0}$

$A_i = -\frac{1}{2}\Sigma_i^{-1}$

$B_i = \Sigma_i^{-1}\mu_i$

$C_{i0} = -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$.

Example: $\begin{bmatrix} 2 \\ 6 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \end{bmatrix} - \omega_1$

$\omega_2 = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} \begin{bmatrix} 3 \\ -4 \end{bmatrix} \begin{bmatrix} 5 \\ 2 \end{bmatrix}$

$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$

$P(\omega_1) = P(\omega_2) = 0.5$.

$\Sigma_1 = \Sigma(x_1 - \mu_1)(x_1 - \mu_1)^T$

$= \begin{bmatrix} -1 \\ 0 \end{bmatrix}_{2\times1} \begin{bmatrix} -1 & 0 \end{bmatrix}_{1\times2}$ $_{2\times1}$

$= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

$\begin{bmatrix} -1 \\ 0 \end{bmatrix}_{2\times1} \times \begin{bmatrix} -1 & 0 \end{bmatrix}_{1\times2} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

$\begin{bmatrix} 0 \\ -2 \end{bmatrix} \times \begin{bmatrix} 0 & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$

$\begin{bmatrix} 0 \\ 2 \end{bmatrix} \times \begin{bmatrix} 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$

$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

$\begin{bmatrix} \cancel{2}/4 \; 1/2 & 0 \\ 0 & \cancel{8}/4 \; 2 \end{bmatrix}$

Logistic Regression:

$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$

$h(x) = \frac{1}{1 + e^{-(\theta x)}}$

$p(x) = \frac{\exp \beta^T x}{1 + \exp \beta^T x}$

$\log\left[\frac{p(x)}{1 - p(x)}\right] = \beta^T x$

$\frac{p(x)}{1 - p(x)} = e^{\beta^T x}$

CS CamScanner

Maximum likelihood estimate:

$$p(x_i) = P(y_i = 1 \mid x_i)$$

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta^T x_i$$

$$L(\beta) = \prod_{y_i=1} p(x_i) \cdot \prod_{y_i=0} (1-p(x_i))$$

$$= \prod_{i=1}^{n} p(x_i)^{y_i} (1-p(x_i))^{1-y_i}$$

$$l(\beta) = \sum_{i=1}^{n} y_i \log p(x_i) + (1-y_i)\log(1-p(x_i))$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right) + \log(1-p(x_i))$$

$$\max \sum_{i=1}^{n} y_i \log(\sigma(x, \beta_1, \beta_0)) + (1-y_i)\log(1-\sigma(x, \beta_1, \beta_0))$$

$$L(y_i, f(x)) = \begin{cases} -\log f(x) & \text{if } y = 1 \\ -\log(1-f(x)) & \text{if } y = 0 \end{cases}$$

$$= y_i \log(f(x)) - (1-y_i)\log(1-f(x))$$

$$\min_{(\beta_1, \beta_0)} \sum_{i=1}^{n} -y_i \log\left(\frac{1}{1+e^{-(\beta_1 x + \beta_0)}}\right) - (1-y_i)\log\left(1 - \frac{1}{1+e^{-(\beta_1 x + \beta_0)}}\right)$$

Gradient Descent for Logistic Regression.

$$\omega_{j+1} = \omega_j - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\left(f(x^{(i)}) - y^{(i)}\right) \cdot x_j^{(i)}\right)$$

$$\beta_1^{(k+1)} = \beta_1^{k} - \eta \cdot \left(-\sum_{i=1}^{m}\left(y_i - \sigma(x, \beta_1, \beta_0)\right) \cdot x_i\right)$$

$$\beta_0^{(k+1)} = \beta_0^{k} - \eta\left(-\sum_{i=1}^{m}\left(y_i - \frac{1}{1+e^{-(\beta_1 x_i + \beta_0)}}\right)\right)$$

Support Vector Machine:

Lagrange multipliers:

~~KKT~~.

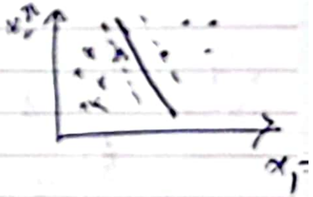
$y_i (w^T x_i + b) \geq 0$

$$\begin{cases} w^T x_i + b > 0 & \omega_1 \\ w^T x_i + b < 0 & \omega_2 \end{cases}$$

Decision boundary $= w^T x + b = 0$

Distance of x from boundary

$$\frac{w^T x + b}{\|w\|}$$



$$\frac{w}{\|w\|}(x_+ - x_-)$$

$$w^T x_+ + b = 1 \; ; \; w^T x_- + b = -1$$

$$w^T (x_+ - x_-) = 2$$

$$\frac{w^T}{\|w\|}(x_+ - x_-) = \frac{2}{\|w\|}$$

Subject to

$$y_i (w^T x_i + b) \geq 1$$

$$L = \frac{1}{2}\|W\|^2 - \sum_{\forall i} \alpha_i [y_i (w^T x + b) - 1]$$

$$\frac{\partial L}{\partial w} = W - \sum_{\forall i} \alpha_i y_i x_i = 0$$

$$\Rightarrow W = \sum_{\forall i} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0$$

$$\Rightarrow \sum \alpha_i y_i = 0$$

$$L = \frac{1}{2} \sum \alpha_i y_i x_i \cdot \sum \alpha_j y_j x_j - \sum \alpha_i y_i x_i \cdot \sum \alpha_j y_j x_j + \sum \alpha_i$$

$$L = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i x_j$$

$\max_{w,b} d$

Equation of a hyperplane:

$$y = mx + b.$$

$$ax + by + c = 0$$

$$y = -\frac{a}{b}x - \frac{c}{b}$$

$$m = \frac{-a}{b}, \quad b = -\frac{c}{b}$$

$$\Rightarrow ax + by + c = 0$$

$$\Rightarrow ax_1 + bx_2 + c = 0$$

$$\Rightarrow W_1 x_1 + W_2 x_2 + W_0 = 0.$$

$$\Rightarrow W^T x + W_0 = 0$$

$$W = [w_1, w_2 \cdots w_n]^T$$

$$X = [x_1, x_2, \cdots, x_n]$$

$$W^T x = 0$$

Hard Margin SVM:

$$\vec{w} \cdot \vec{u} \geq c$$

$$\vec{w} \cdot \vec{u} - c \geq 0$$

$$\vec{w} \cdot \vec{u} + b \geq 0 \quad \{+\text{ive if } <0 \ -\text{ive}\}$$

$$\hat{y} = \begin{cases} +1 & \text{if } \vec{w} \cdot \vec{x}_i + b \geq 0 \\ -1 & \text{if } \vec{w} \cdot \vec{x}_i + b < 0 \end{cases}$$

$$W^T x + b = 1 \ \& \ W^T x + b = -1$$

$$\begin{cases} \vec{w} \cdot \vec{x} + b \geq 1 \\ \vec{w} \cdot \vec{x} + b \leq -1 \end{cases} = \begin{cases} y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1 \\ y_i (\vec{w} \cdot \vec{x} + b) \geq 1 \end{cases}$$

$$y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1. \quad \{\text{for supp vecs. } = 1\}$$

$$d = (\vec{x}_2 - \vec{x}_1) \cdot \frac{\vec{w}}{\|W\|}$$

$$\Rightarrow d = \frac{\vec{x}_2 \cdot \vec{w} - x_1 \cdot \vec{w}}{\|W\|}$$

$$\Rightarrow d = y_i (\vec{w} x_i + b) = 1 \Rightarrow 1(\vec{w} \cdot x_2 + b) = 1$$

$$\Rightarrow \vec{w} \cdot x_2 = 1 - b.$$

$$\Rightarrow \frac{1 - b - (-b - 1)}{\|W\|} = \frac{1 - \cancel{b} + \cancel{b} + 1}{\|W\|}$$

$$\Rightarrow d = \frac{2}{\|W\|}$$

$$\arg\max_{w,b} = \frac{2}{\|W\|} \quad \text{such that: } \{ y_i (\vec{w} \cdot \vec{x}_i + b \geq$$

Soft Margin

max $f(x) \Leftrightarrow$ min $\frac{1}{f(x)}$

$$\underset{(w,b)}{\arg\min} = \frac{\|W\|}{2} + \sum_{i=1}^{n} \xi_i$$

ξ_i for correctly classified pt.

S.V.M error.

Margin error + classification error.

$$\underset{(w,b)}{\arg\min} \; \frac{\|W\|}{2} + C\sum_{i=1}^{n} \xi_i$$

If we increase C it focuses on classification error.

$C \propto 1/\lambda$

P.C.A.

For $R^n \to R^1$

$x_i \in R^n, u \in R^n$; $\bar{u}$ = Mean

$$\text{Variance} = \frac{1}{l-1}\sum_{i=1}^{l}(u^T x_i - u^T \bar{x}_i)^2$$

$$= \frac{1}{l-1}\sum_{i=1}^{l}(u^T(x_i - \bar{u}))^2$$

$$= \frac{1}{l-1}\sum_{i=1}^{l}(u^T(x_i-\bar{u})(x_i-\bar{u})^T u)$$

$$= \frac{1}{l-1}\sum_{i=1}^{l}(u^T \Sigma u)$$

$x_i \in R^n$

$(u^T x_i)u$

$$\text{Min} \sum_{i=1}^{l}(x_i - (u^T x_i)u)^2$$

$$= \sum_{i=1}^{l}(x_i - (u^T x_i)u)^T(x_i - (u^T x_i)u)$$

$$= \sum_{i=1}^{l} x_i^T x + ((u^T x_i)u)^T((u^T x_i)u) - 2x_i^T(u^T x_i)u$$

$$= \sum_{i=1}^{l} x_i^T x_i + (u^T x_i)(x_i^T u)$$

$$= \frac{1}{l}\sum_{i=1}^{l} - u^T\left(\frac{1}{n}\sum_{i=1}^{l}(x_i x_i^T)\right)u$$

$\underset{u \in R^n}{\text{Max}}\; u^T \Sigma u$. STC = $u^T u = 1$.

$\underset{u \in R^n}{\text{Min}}\; -u^T \Sigma u$ STC = $u^T u = 1$

$L(u,\alpha) = -u^T \Sigma u + \alpha_i(u^T u - 1)$

$\nabla_u L(u,\alpha) = -\Sigma u + \alpha u = 0$

$\Rightarrow \Sigma u = \alpha u$ At pt. of optimality

$u^T \alpha u = \alpha u^T u = \alpha$ (Max value of eigen value of Σ

$$\frac{d}{du}(Y - Au)^T \cdot (Y - Au)$$

$$A^T(Y - Au) + (Y - Au)^T \times A.$$

$$A^T(Y - Au) + A^T(Y - Au)$$

$$2\left(A^T(Y - Au)\right) = 0$$

$$A^T(Y - Au) = 0$$

$$\Rightarrow A^T Y - A^T A u = 0$$

$$\Rightarrow A^T Y = A^T A u$$

$$\Rightarrow (A^T A)^{-1} A^T Y = (A^T A)^{-1} A^T A u.$$

$$(A^T A)^{-1} A^T Y.$$

MSE

Sc