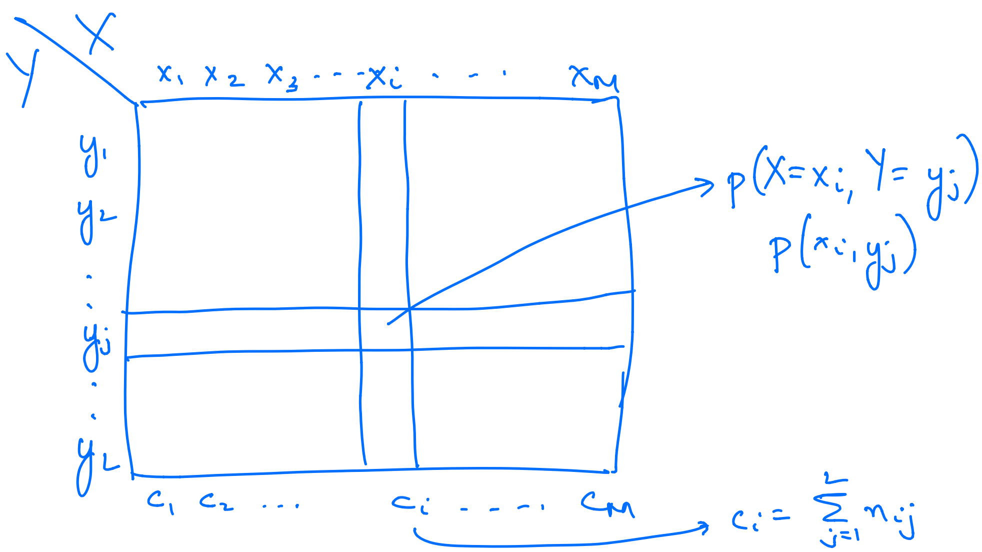
Probability Theory Background

Random Variables X & Y

X: $x_1, x_2, \dots, x_M$

Y: $y_1, y_2, \dots, y_L$

Joint Distribution $p(X, Y)$



N trials, let n_{ij} be the no. of times we observe $X = x_i$ & $Y = y_j$.

As $N \to \infty$, $p(x_i, y_j) = \frac{n_{ij}}{N}$

marginal distribution

$$\boxed{p(X=x_i) = \sum_{j=1}^{L} p(X=x_i, Y=y_j)} = \sum_{j=1}^{L} \frac{n_{ij}}{N} = \frac{c_i}{N}$$

Sum Rule

Conditional Probability of $Y = y_j$ given $X = x_i$

$$p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

$$p(y_j \mid x_i)$$

$$p(y_j, x_i) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

Product Rule

$$\boxed{p(Y = y_j, X = x_i) = p(Y = y_j \mid X = x_i) \cdot p(X = x_i)}$$

Sum Rule $\boxed{p(X) = \sum_Y p(X, Y)}$

Product Rule $\boxed{p(X, Y) = p(Y \mid X)\, p(X)}$

$$p(Y \mid X)\, p(X) = p(X, Y) = p(X \mid Y)\, p(Y)$$

Bayes Rule, $p(Y \mid X) = \dfrac{p(X \mid Y)\, p(Y)}{p(X)}$ → prior

posterior

evidence
"data-likelihood"

$$= \frac{p(X \mid Y)\, p(Y)}{\sum_Y p(X, Y)}$$

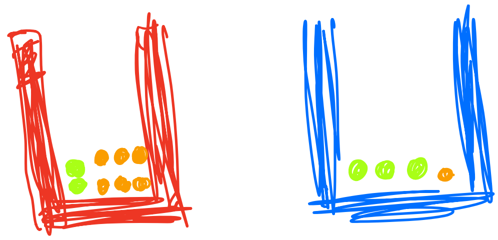
$$= \frac{p(X \mid Y)\, p(Y)}{\sum_Y p(X \mid Y)\, p(Y)}$$

Independence : $p(x_i, y_j) = p(x_i)\,p(y_j)$, $p(y_j | x_i) = p(y_j)$

2a & 6o   3a & 1o

Two Boxes : Red & Blue

Two kinds of fruit : Apples & Oranges



$p(r) = 0.4 = \frac{2}{5}$

$p(b) = 0.6 = \frac{3}{5}$

$p(a|r) = \frac{1}{4}$, $p(o|r) = \frac{3}{4}$

$p(a|b) = \frac{3}{4}$, $p(o|b) = \frac{1}{4}$

F \ B	r	b	
a	1/10	9/20	11/20
o	3/10	3/20	9/20
	2/5	3/5	

$p(a, r) = p(a|r)\,p(r) = \frac{1}{4} \cdot \frac{2}{5} = \frac{1}{10}$

$p(o, r) = p(o|r)\,p(r) = \frac{3}{4} \cdot \frac{2}{5} = \frac{3}{10}$

$p(a, b) = p(a|b)\,p(b) = \frac{3}{4} \cdot \frac{3}{5} = \frac{9}{20}$

$p(o, b) = p(o|b)\,p(b) = \frac{1}{4} \cdot \frac{3}{5} = \frac{3}{20}$

$p(a) = p(a, r) + p(a, b) = \frac{1}{10} + \frac{9}{20} = \frac{11}{20}$

$$p(r|o) = \frac{p(o|r)\,p(r)}{p(o)} = \frac{p(o, r)}{p(o)} = \frac{3/10}{9/20} = \frac{3}{10} \cdot \frac{20}{9} = \frac{2}{3}$$

Probabilities wrt continuous variables, $x \in \mathbb{R}$, $x \in \mathbb{R}^d$

pdf = probability density function $p(x)$

$p(x) \geq 0$, $\int_{-\infty}^{\infty} p(x)\,dx = 1$

Sum Rule : $p(x) = \int_{-\infty}^{\infty} p(x,y)\,dy$

Product Rule: $p(x,y) = p(y|x)p(x) = p(x|y)p(y)$

Expectation (Mean)

$$E[f(x)] = \sum_x p(x)f(x) \qquad \int p(x)f(x)\,dx$$

Variance

$(E[x+y] = E(x) + E[y])$

$$\begin{aligned}
Var[f(x)] &= E\left[\left(f(x) - E[f(x)]\right)^2\right] \\
&= E\left[(f(x))^2 + (E[f(x)])^2 - 2f(x)E[f(x)]\right] \\
&= E[(f(x))^2] + E[(E[f(x)])^2] - 2E[f(x)E[f(x)]] \\
&= E[(f(x))]^2 - (E[f(x)])^2
\end{aligned}$$

$$Var(x) = E[(x - E[x])^2] = E[x^2] - (E[x])^2$$

$$\begin{aligned}
cov(x,y) &= E[\{x - E[x]\}\{y - E[y]\}] \\
&= E[xy] - E[x]E[y]
\end{aligned}$$

What if x & y are independent? $p(x,y) = p(x)p(y)$

$$cov(x,y) = E[xy] - E[x]E(y)$$

$$\int p(x,y)\,xy\,dx\,dy = \int p(x)p(y)\,x\,y\,dx\,dy$$

$$= \int p(x)x\,dx \cdot \int p(y)\cdot y\cdot dy$$

$$= E[x]E[y]$$

$cov(x,y) = 0$ if x & y are independent.

Gaussian Distribution / Normal Distribution

$x \in \mathbb{R}$

$$p(x \mid \mu, \sigma^2) = p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{1}{2}(x-\mu)^2/\sigma^2},$$

μ = mean or expectation

σ^2 = variance

$$\int_{-\infty}^{\infty} x\,p(x)\,dx = \mu = E[x]$$

$$E[(x-\mu)^2] = E[x^2] - \mu^2 = \sigma^2$$

$x \in \mathbb{R}^d$, $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ $\quad E[x] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \mu$, $E[(x_i-\mu_i)^2] = \sigma_i^2$

$$p(x) = p(x_1, x_2, \ldots x_d)$$

Suppose x_i is indendent of x_j, $\forall\, i \neq j$

$$p(x) = \prod_{i=1}^{d} p(x_i)$$

$$= \frac{1}{(\sqrt{2\pi})^d \sigma_1 \sigma_2 \ldots \sigma_d} \prod_{i=1}^{d} e^{-\frac{1}{2}(x_i-\mu_i)^2/\sigma_i^2}$$

$$= \frac{1}{(2\pi)^{d/2} \underbrace{\sigma_1\sigma_2\ldots\sigma_d}_{\to \det(\Sigma)^{1/2}}} \; e^{-\frac{1}{2}\sum_i (x_i-\mu_i)^2/\sigma_i^2} \to e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

$$(x-\mu)^T(x-\mu) = \sum_i (x_i-\mu_i)^2$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_i (x_i-\mu_i)^2/\sigma_i^2 \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_d^2 \end{bmatrix}$$

General Case : $x \in \mathbb{R}^d$

Multivariate Gaussian Distribution :

$$p(x) = \frac{1}{(2\pi)^{d/2}\, \underbrace{|\Sigma|}_{\text{determinant of } \Sigma}^{1/2}}\, e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Σ = Covariance Matrix = $E[(x-\mu)(x-\mu)^T]$

Σ_{ij} is covariance between x_i & x_j

Σ is $d\times d$, symmetric, positive definite

$\Sigma = V\Lambda V^T$ (Λ is diagonal, $\Lambda_{ii} > 0$
$V^TV = I$, $VV^T = I$)

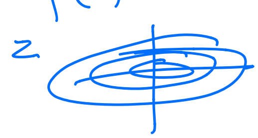
$\Sigma^{-1} = V\Lambda^{-1}V^T$

$$\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = \frac{1}{2}(x-\mu)^T V\Lambda^{-1}V^T(x-\mu)$$

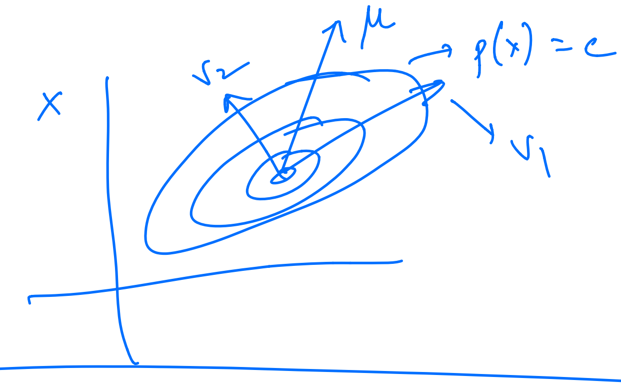
$$z = V^T(x-\mu)$$

$$= \frac{1}{2} z^T \Lambda^{-1} z$$

$p(x) = c \Rightarrow \frac{1}{2} z^T\Lambda^{-1} z = c$



$$\boxed{\frac{1}{2}\sum \frac{z_i^2}{\sigma_i^2} = c} \rightarrow \text{Equation of ellipse}$$



Singular Value Decomposition (SVD)

$A \in \mathbb{R}^{m\times n}$

$A = U \Sigma V^T$

singular values

$\hat{\Sigma}$

$$\underbrace{\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}}_{m \times n} = \underbrace{\begin{bmatrix} u_1 & u_2 & \cdots & u_n & u_{n+1} & \cdots & u_m \end{bmatrix}}_{U,\ m\times m;\ \hat{U} = [u_1 \cdots u_n]} \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 \ddots & \\ 0 & & \sigma_n \\ \hline & 0 & \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}}_{V^T,\ n \times n}$$

left singular vectors

right singular vectors

$UU^T = U^TU = I$, $\quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$, $\quad VV^T = V^TV = I$

$A = U \Sigma V^T \qquad (U \in \mathbb{R}^{m\times m},\ \Sigma \in \mathbb{R}^{m \times n},\ V \in \mathbb{R}^{n\times n})$

$AV = U\Sigma$

$Av_i = u_i \sigma_i$

"Thin or reduced SVD": $A = \hat{U}\hat{\Sigma}\hat{V}^T$, $\quad \hat{U} \in \mathbb{R}^{m\times n}$, $\hat{\Sigma} \in \mathbb{R}^{n\times n}$, $\hat{V} \in \mathbb{R}^{n\times n}$

$m \geq n$ diagonal matrix

$$\hat{U}^T\hat{U} = I \quad , \quad \hat{U}\hat{U}^T \neq I$$

$$\hat{V}^T\hat{V} = I$$

$$n \quad \overset{n}{A^TA} = (\hat{V}\hat{\Sigma}\hat{U}^T)(\hat{U}\hat{\Sigma}\hat{V}^T) = \hat{V}\hat{\Sigma}^2\hat{V}^T \longrightarrow \text{eigenvalue decomposition of } A^TA$$

$$m \quad \overset{m}{AA^T} = (\hat{U}\hat{\Sigma}\hat{V}^T)(\hat{V}\hat{\Sigma}\hat{U}^T) = \hat{U}\hat{\Sigma}^2\hat{U}^T$$

$$A = U\Sigma V^T \quad , \quad A^T = V\Sigma^T U^T$$

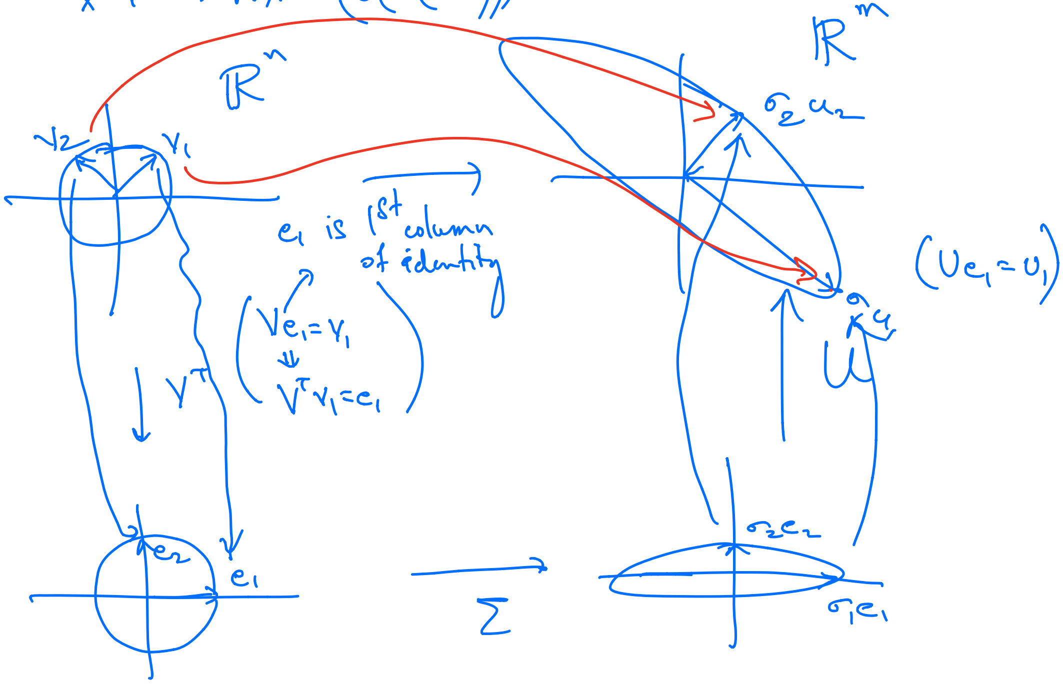
$$AV = U\Sigma \qquad A^T U = V\Sigma^T$$

$$A v_i = u_i\sigma_i \qquad A^T u_i = v_i\sigma_i \quad , \quad i = 1, 2, \ldots n$$

$$A^T u_i = 0 \quad , \quad i = n+1, \ldots m$$

$$A : \mathbb{R}^n \to \mathbb{R}^m$$

$$x \longmapsto Ax = (U(\Sigma(V^T x)))$$



If A has rank r

$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0, \quad \sigma_{r+1} = \sigma_{r+2} \dots = \sigma_n = 0$

$A : \mathbb{R}^n \to \mathbb{R}^m$

$$\begin{aligned}
v_1 &\mapsto u_1\sigma_1 \\
v_2 &\mapsto u_2\sigma_2 \\
&\vdots \\
v_r &\mapsto u_r\sigma_r \\
v_{r+1} &\mapsto 0 \\
&\vdots \\
v_n &\mapsto 0
\end{aligned}$$

$A^T : \mathbb{R}^m \to \mathbb{R}^n$

$$\begin{aligned}
u_1 &\mapsto v_1\sigma_1 \\
u_2 &\mapsto v_2\sigma_2 \\
&\vdots \\
u_r &\mapsto v_r\sigma_r \\
u_{r+1} &\mapsto 0 \\
&\vdots \\
u_m &\mapsto 0
\end{aligned}$$

SVD provides orthogonal basis for the four fundamental subspaces of A

Column Space $= R(A) = \langle u_1, u_2 \dots u_r \rangle$

Row Space $= R(A^T) = \langle v_1, v_2, \dots v_r \rangle$

Null Space$(A) = N(A) = \langle v_{r+1}, \dots v_n \rangle$

Null Space$(A^T) = N(A^T) = \langle u_{r+1}, \dots\dots, u_m \rangle$

$A \in \mathbb{R}^{m \times n}$

Truncated SVD, $A_k = U_k \Sigma_k V_k^T$, $U_k \in \mathbb{R}^{m \times k}$, $\Sigma_k \in \mathbb{R}^{k \times k}$, $V_k \in \mathbb{R}^{n \times k}$

$U_k^T U_k = I$

$V_k^T V_k = I$

Among all rank-k approximations of A, A_k is the "best"

$$A_k = \arg\min_{B \text{ of rank } k} \|A - B\|_2 \quad , \quad A_k = \arg\min_{B \text{ of rank } k} \|A - B\|_F$$

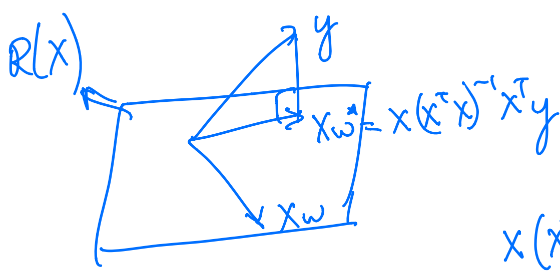
Regression

$$\boxed{\min_w \|y - Xw\|_2^2} \quad , \quad X = \begin{array}{cc} & d+1 \\ N & \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \\ 1 & x_N^T \end{bmatrix} \end{array}$$

Least Squares Solution $\quad X^TXw^* = X^Ty$

$$w^* = (X^TX)^{-1}X^Ty$$

Prediction on training set : $Xw^* = X(X^TX)^{-1}X^Ty$



$U \in \mathbb{R}^{N \times (d+1)}$

$X = U\Sigma V^T$ — reduced SVD

$X(X^TX)^{-1}X^T \rightarrow$ "hat" matrix

$$(U\Sigma V^T)(V\Sigma \underbrace{U^TU}_{I}\Sigma V^T)^{-1} V\Sigma U^T$$

$$U\Sigma V^T(V\Sigma^2V^T)^{-1} V\Sigma U^T$$

$$U\Sigma \underbrace{V^T(V}_{I}\Sigma^{-2}\underbrace{V^T)V}_{I}\Sigma U^T = UU^T$$

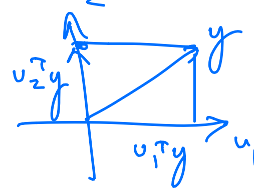
$U = [u_1\ u_2 \dots u_{d+1}]$

$U^TU = I \quad , \quad UU^T \neq I$ (orthogonal projector)

$$UU^T = [u_1\ u_2 \dots u_{d+1}] \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_{d+1}^T \end{bmatrix} = u_1u_1^T + u_2u_2^T + \dots + u_{d+1}u_{d+1}^T$$

Least Squares Prediction $= \boxed{UU^T} y \to \sum u_i u_i^T$

$$= \sum_{i=1}^{n} u_i (u_i^T y)$$



Least Squares Regression: $\min_w \|y - Xw\|_2^2$

Ridge Regression: $\min_w \|y - Xw\|_2^2 + \lambda \|w\|_2^2 \quad , \lambda \geq 0$

Solution: $(X^T X + \lambda I) w^* = X^T y$

$$\Rightarrow \quad w^* = (X^T X + \lambda I)^{-1} X^T y$$

$N \geq d+1$

Prediction: $X w^* = X (X^T X + \lambda I)^{-1} X^T y$

$X = U \Sigma V^T$ — reduced SVD

$X^T X = V \Sigma^2 V^T$

$X^T X + \lambda I = V(\Sigma^2 + \lambda I) V^T \qquad (VV^T = I)$

$(X^T X + \lambda I)^{-1} = V (\Sigma^2 + \lambda I)^{-1} V^T$

$$U \Sigma \underbrace{V^T \cdot V}_{I} (\Sigma^2 + \lambda I)^{-1} \underbrace{V^T V}_{I} \Sigma U^T$$

$$U \underbrace{\Sigma (\Sigma^2 + \lambda I)^{-1} \Sigma}_{} U^T$$

$$\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2 + \lambda} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2 + \lambda} & & \\ & & \ddots & \\ & & & \frac{\sigma_{d+1}^2}{\sigma_{d+1}^2 + \lambda} \end{bmatrix}$$

Ridge Regression Solution = $\sum u_i \left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right) v_i^T y$

Shrinkage

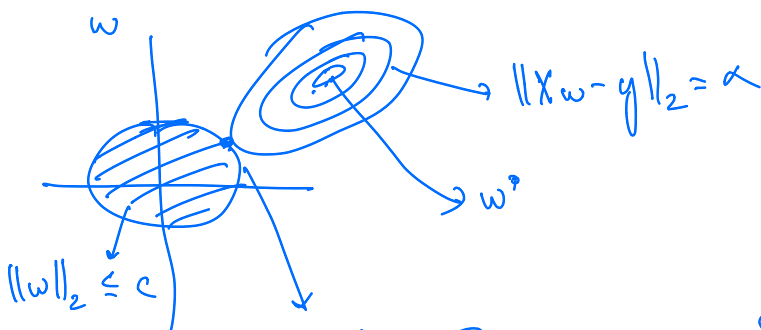
$\sigma_i^2 >> \lambda$, $\frac{\sigma_i^2}{\sigma_i^2+\lambda} \simeq 1$

$\sigma_i^2 << \lambda$, $\frac{\sigma_i^2}{\sigma_i^2+\lambda} \simeq 0$

Ridge Regression can equivalently be thought of as:

$$\min_{w} \|Xw-y\|_2$$

$$\text{st} \quad \|w\|_2 \leq c$$



<u>Ridge Regression Solution</u>