CS761 Notes

Young Wu

December 14, 2018

1 Lecture 1

1.1 CLASSIFY: like GG or NOT

$$y = \begin{cases} 1 & \text{if like GG} \\ 0 & \text{if not} \end{cases}$$

$x_1 =$ number of stars for SS
$x_2 =$ number of stars for CRA

$$\hat{y} = \begin{cases} +1 & \text{if } \dfrac{\hat{\mathbb{P}}\left\{y = 1 | x_1, x_2\right\}}{\hat{\mathbb{P}}\left\{y = -1 | x_1, x_2\right\}} \geqslant 1 \\ -1 & \text{otherwise} \end{cases}$$

1.2 Random variables

X and Y are random variables
Joint prob distribution: $p(x, y) = \mathbb{P}\left\{X = x \text{ and } Y = y\right\}$
Marginal: $p(x) = \displaystyle\sum_{y \in \text{ all possible } y \text{ values}} p(x, y)$
Conditional dist: $p(y|x) = \dfrac{p(x, y)}{p(x)}$
Bayes Rule: $p(y|x) = \dfrac{p(x|y)\, p(y)}{p(x)}$
Expectation: $\mu = \mathbb{E}[X] = \displaystyle\sum_x x p(x), \hat{\mu} = \dfrac{1}{40} \sum_{i=1}^{40} x_i$

$$\begin{aligned}
\mathbb{E}[f(X)] &= \sum_x f(x)\, p(x) \\
\mathbb{E}\left[X^2\right] &= \sum_x x^2 p(x) \\
\mathbb{E}[X + Y] &= \sum_x \sum_y (x + y)\, p(x, y) \\
&= \sum_x x \sum_y p(x, y) + \sum_y y \sum_x p(x, y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y] \\
\mathbb{E}[XY] &= \sum_x \sum_y xy p(x, y)
\end{aligned}$$

Independence iff $p(x,y) = p(x)\,p(y)$
If $X \perp Y$,

$$\begin{aligned}\mathbb{E}[XY] &= \sum_x \sum_y xyp(x)\,p(y)\\ &= \sum_x xp(x) \sum_y yp(y)\\ &= \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

Variance:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]\\ &= \sum_x (x - \mathbb{E}[X])^2\, p(x)\end{aligned}$$

Conditional Expectation: $\mathbb{E}[Y|X=x] = \sum_y yp(y|x)$
Sums of Independent random variables
$x_1, x_2, ...$ are indep random variables
$S_n = \sum_{i=1}^n x_i$, what is distribution of S_n

$$\begin{aligned}\mathbb{E}[S_n] &= \sum_{i=1}^n \mathbb{E}[X_i]\\ \mathbb{V}[S_n] &= \mathbb{E}\left[(S_n - \mathbb{E}[S_n])^2\right]\\ &= \mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)^2\right]\\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i])\,(X_i - \mathbb{E}[X_i])\right]\\ &= \sum_{i=1}^n \mathbb{E}\left[(X_i - E[X_i])^2\right]\\ &= \sum_{i=1}^n \mathbb{V}[X_i]\\ \mathbb{E}[X_i - \mathbb{E}[X_i]] &= 0\end{aligned}$$

1.3 Example

$$\begin{aligned}p &= \mathbb{P}\{\text{ unif randomly chosen student SS} = 3,\ \text{CRA} = 4\ \}\\ \hat{p} &= \frac{1}{40}\sum_{i=1}^{40} \mathbb{1}_{\text{ ith person says SS} = 3,\ \text{CRA} = 4}\\ &= \frac{1}{40}\sum_{i=1}^{40} \mathbb{1}_{i,3,4}\end{aligned}$$

$$\mathbb{E}\left[\hat{p}\right] = \frac{1}{40}\sum_{i=1}^{40}\mathbb{E}\left[\mathbb{1}_{i,3,4}\right] = p$$

Unbiased:

$$\begin{aligned}
\mathbb{V}\left[\hat{p}\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{i,3,4} - p\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\left(\mathbb{1}_{i,3,4} - p\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbb{1}_{i,3,4} - p\right)^2\right]
\end{aligned}$$

and,

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbb{1}_{i,3,4} - p\right)^2\right] &= (1-p)^2 p + p^2(1-p) = p(1-p) \\
\mathbb{E}\left[\hat{p}\right] &= p \\
\mathbb{V}\left[\hat{p}\right] &= \frac{p(1-p)}{n} \\
\text{std }(\hat{p}) &= \sqrt{\frac{p(1-p)}{n}}
\end{aligned}$$

2 Lecture 2

2.1 Discrete Random Variables

Y random variable taking values in $\{a_1, ..., a_m\}$.

$$\begin{aligned}
p_j &= \mathbb{P}\left\{Y = a_j\right\}, j = 1, ..., m \\
\sum_{j=1}^{m} p_j &= 1
\end{aligned}$$

Bernoulli:

$$\begin{aligned}
Y &\in \{0, 1\} \\
p &= \mathbb{P}\left\{Y = 1\right\} \\
\mathbb{P}\left\{Y = y\right\} &= p^y(1-p)^{1-y} \\
\mathbb{E}\left[Y\right] &= 1\cdot p + 0\cdot(1-p) \\
\mathbb{V}\left[Y\right] &= \mathbb{E}\left[(Y-p)^2\right] = p(1-p)
\end{aligned}$$

Binomial:

$$\begin{aligned}
Y_1, Y_2, ..., Y_n &\overset{iid}{\sim} \text{ Be }(p) \\
\mathbb{P}\left\{Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n\right\} &= \prod_{i\hat{=}1n}\mathbb{P}\left\{Y_i = y_i\right\}
\end{aligned}$$

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

Binomial random variable with params n, p

$$K := \sum_{i=1}^{n} Y_i \sim \text{ Bi } (n, p)$$
$$\mathbb{P}\{K = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

Bin coef: $\dfrac{n!}{k!\,(n-k)!}$

Multinomial

$$Y \in \{a_1, ..., a_m\}, i = 1, ...n, \text{ indep}$$
$$\mathbb{P}\{Y_1 = y_1, ..., Y_n = y_n\} = \prod_{i=1}^{n} \mathbb{P}\{Y_i = y_i\}$$
$$= \prod_{i=1}^{n} \prod_{i=1}^{m} p_j^{\mathbb{1}_{y_i = j}}$$
$$K_j = \{\text{ number times } a_j \text{ appears in } Y_1, ...Y_n\}$$
$$\mathbb{P}\{K_1 = k_1, ..., K_m = k_m\} = \underbrace{\binom{n}{k_1, ..., k_m}}_{\text{multinomial coef}} \prod_{j=1}^{m} p_j^{k_j}$$

Poisson

$$X \geqslant 0 \text{ integer-valued}$$
$$\mathbb{P}\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \lambda > 0 \text{ param}$$
$$\mathbb{E}[X] = \lambda$$
$$\mathbb{V}[X] = \lambda$$

2.2 Optimal Binary Classification

feature X, label $Y \in \{0, 1\}$

$$(X, Y) \sim \mathbb{P}_{XY}$$
$$\{(X_i, Y_i)\}_{i=1}^{n} \overset{iid}{\sim} \mathbb{P}_{XY}$$

Classifier

$$f : X \to Y$$
$$\hat{y} = f(X)$$

Loss: 0/1 loss

$$Loss\left(\hat{y}, y\right) = \mathbb{1}_{\hat{y} \neq y} = \begin{cases} 1 & \text{if } \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases}$$

Risk: expected loss

$$R\left(f\right) = \mathbb{E}\left[\mathbb{1}_{f(X) \neq Y}\right] = \mathbb{P}\left\{f\left(X\right) \neq Y\right\}$$

Bayes Risk:

$$R^{\star} = \inf_{f} R\left(f\right)$$

min probability of error

Bayes Optimal Classifier

$$f^{\star}\left(x\right) = \mathbb{1}_{\underbrace{\mathbb{P}\left\{Y = 1 | X = x\right\}}_{\eta(x)} = \frac{1}{2}}$$

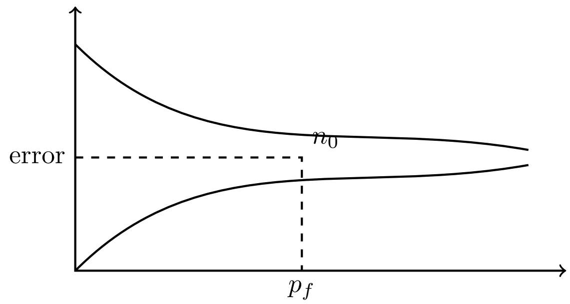
pick label that is most probable given x

$$\begin{aligned} R^{\star} &= \mathbb{E}\left[\min\left\{\eta\left(x\right), 1 - \eta\left(x\right)\right\}\right] \\ \hat{y} = 1 &\rightarrow p\left(\text{ err }\right)) = 1 - \eta\left(x\right) \\ \hat{y} = 0 &\rightarrow p\left(\text{ err }\right)) = \eta\left(x\right) \end{aligned}$$

Theorem 1. $R\left(f^{\star}\right) = R^{\star}$

Estimating $\mathbb{P}\left\{f\left(X\right) \neq Y\right\} =: p_f$
labeled examples $\left\{\left(X_i, Y_i\right)\right\}_{i=1}^{n} \overset{iid}{\sim} \mathbb{P}_{XY}$

$$\begin{aligned} \hat{p}_f &= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{1}_{f(X_i) \neq Y_i}}_{\text{iid Bernoulli random variable}} \\ \mathbb{E}\left[\hat{p}_f\right] &= p_f \\ \mathbb{E}\left[\left|n\hat{p}_f - np_f\right|^2\right] &= np_f\left(1 - p_f\right) \\ \mathbb{E}\left[\left|\hat{p}_f - p_f\right|^2\right] &= \frac{p_f\left(1 - p_f\right)}{n} = \sigma^2 \\ \sigma &= \sqrt{\frac{p_f\left(1 - p_f\right)}{n}} \end{aligned}$$



2.3 Analysis of Nearest Neighbor Classifier

Given $\{(X_i, Y_i)\}_{i=1}^{n}$

$$X_i \in \mathbb{R}^d$$

predict Y for new X

Theorem 2. *(Cover and Hart* $60's$*)*$\mathbb{E}\left[R_n\left(X\right)\right] =$ *expected error of NN classifier*

$$\lim_{n\to\infty} \mathbb{E}\left[R_n\left(x\right)\right] = \mathbb{E}\left[2\eta\left(x\right)\left(1-\eta\left(x\right)\right)\right] \leqslant 2R^{\star}$$

as $n \to \infty$, NN of X, say $x' \in \{x_i\}_{i=1}^{n}, x \approx x'$
errs: $Y = 1, Y' = 0$ or $Y = 0, Y' = 1$

$$\eta\left(x\right)\left(1-\eta\left(x'\right)\right) + \left(1-\eta\left(x\right)\right)\eta\left(x'\right) \approx 2\eta\left(x\right)\left(1-\eta\left(x\right)\right)$$

$$\begin{aligned}
z :&= \min\left\{\eta\left(x\right), 1-\eta\left(x\right)\right\} \\
\eta\left(x\right)\left(1-\eta\left(x\right)\right) &= z\left(1-z\right) \\
\mathbb{E}\left[\eta\left(x\right)\left(1-\eta\left(x\right)\right)\right] &= \mathbb{E}\left[z\left(1-z\right)\right] = \mathbb{E}\left[z-z^2\right] \\
&= \mathbb{E}\left[z\right] - \underbrace{\mathbb{E}\left[z^2\right]}_{\geqslant \mathbb{E}[z]^2 \text{ Jensen's Inequality}} \\
&\leqslant \mathbb{E}\left[z\right] - \mathbb{E}\left[z\right]^2 \\
&= \mathbb{E}\left[z\right]\left(1-\mathbb{E}\left[z\right]\right) \\
&\leqslant \mathbb{E}\left[z\right]
\end{aligned}$$

3 Lecture 3

3.1 Matrices

$$\begin{aligned}
&u_1, u_2, u_3 \\
&\|u_1\| = \|u_2\| = 1 \\
&u_i^T u_j = 0, i \neq j
\end{aligned}$$

$$U = \left[u_1 u_2 ... u_m \right], u_i \in \mathbb{R}^n$$

Orthogonal matrix

$$\begin{aligned}
U^T U &= I_m \\
U^T U &= \begin{bmatrix} u_1^T \\ ... \\ u_m^T \end{bmatrix} \begin{bmatrix} u_1 & ... & u_m \end{bmatrix} \begin{bmatrix} u_1^T u_1 & ... & u_1^T u_m \\ ... & ... & ... \\ u_m^T u_1 & ... & u_m^T u_m \end{bmatrix} \\
&= \begin{bmatrix} 1 & ... & 0 \\ ... & ... & ... \\ 0 & ... & 1 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\mathbb{R}^n &= S \oplus S^\perp \\
&\{\hat{u}_1, \hat{u}_2\} \in S, \{\hat{u}_3\} \in S^\perp
\end{aligned}$$

$$y = Ax, y \in \mathbb{R}^m, A \in M_{m \times n}, x \in \mathbb{R}^n$$

$$\begin{aligned}
\begin{bmatrix} y_r \\ y_m \end{bmatrix} &= \begin{bmatrix} \sigma_i & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_r \\ x_m \end{bmatrix} \\
y_r &= \begin{bmatrix} y_1 \\ ... \\ y_r \end{bmatrix} \\
y_m &= \begin{bmatrix} y_{r+1} \\ ... \\ y_m \end{bmatrix}, \text{ always zero} \\
\sigma_i &= \begin{bmatrix} \sigma_1 & ... & 0 \\ ... & ... & ... \\ 0 & ... & \sigma_r \end{bmatrix} \\
x_r &= \begin{bmatrix} x_1 \\ ... \\ x_r \end{bmatrix} \\
x_m &= \begin{bmatrix} x_{r+1} \\ ... \\ x_n \end{bmatrix}, \text{ don't matter}
\end{aligned}$$

$$\mathbb{R}^m = \underbrace{N(A)}_{n-r} \oplus \underbrace{N(A)^\perp}_{r}$$

where $N(A)$ is the nullspace

$$\mathbb{R}^m = \underbrace{R(A)}_{r} \oplus \underbrace{R(A)^{\perp}}_{n-r}$$

3.2 SVD

$$\begin{aligned}
\underbrace{A}_{m\times n} &= \underbrace{U}_{m\times m}\underbrace{\Sigma}_{m\times n}\underbrace{V^T}_{n\times n} \\
U^T U &= I
\end{aligned}$$

if U square, $U^T = U^{-1}$

$$\begin{aligned}
\|Ux\|^2 &= (Ux)^T (Ux) = x^T U^T U x = x^T x = \|x\| \\
(Ux)^T (Uy) &= x^T U^T U y = x^T y
\end{aligned}$$

$$\begin{aligned}
y &= Ax = U\Sigma V^T x, V^T x = \begin{bmatrix} v_1^T x \\ ... \\ v_n^T x \end{bmatrix} \\
VV^T x &= \left(v_1 v_1^T + v_2 v_2^T + v_3 v_3^T\right) x = \left(v_1^T x\right) v_1 + \left(v_2^T x\right) v_2 + \left(v_3^T x\right) v_3 \\
x &= v_i \\
y &= \sigma_i u_i \\
r &= \text{ rank } (A)
\end{aligned}$$

$$\begin{aligned}
R(A) &= \{u_1, ..., u_r\} \\
R(A)^{\perp} &= \{u_{r+1}, ..., u_m\} \\
N(A) &= \{v_{r+1}, ..., r_n\} \\
N(A)^{\perp} &= \{v_1, ..., v_r\}
\end{aligned}$$

3.3 Matrix Identities

$$\begin{aligned}
\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}^{-1} &= \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \\
\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\begin{bmatrix} X \\ Y \end{bmatrix} &\sim \left(0, \begin{bmatrix} \Sigma_x & \Sigma_{x,y} \\ \Sigma_{y,x} & \Sigma_y \end{bmatrix}\right) \\
X|Y &\sim \left(..., \Sigma_x - \Sigma_{x,y}\Sigma_{y^{-1}}\Sigma_{y,x}\right)
\end{aligned}$$

$$
\begin{aligned}
Ax_1 + Bx_2 &= y_1 \\
Cx_1 + Dx_2 &= y_2 \\
x_1 &= A^{-1}\left(y_1 - Bx_2\right) \\
CA^{-1}y_1 - CA^{-1}Bx_2 + Dx_2 &= y_2 \\
\left(D - CA^{-1}B\right)x_2 &= \left(y_2 - CA^{-1}y_1\right)
\end{aligned}
$$

Matrix Iversion Lemma

$$
\left(A - BD^{-1}C\right)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1}
$$

Other Identities:

$$
\begin{aligned}
A\left(I + A\right)^{-1} &= I - \left(I + A\right)^{-1} \\
A &= \left(I + A\right) - I
\end{aligned}
$$

3.4 Vector derivatives

$$
\begin{aligned}
f\left(x\right) &= 0, f : \mathbb{R}^n \to \mathbb{R} \\
\frac{\partial f}{\partial x_1} &= 0 \text{ and } \frac{\partial f}{\partial x_2} = 0 \\
\frac{df}{dx} &= \text{ gradient } = \nabla f \\
f &: \mathbb{R}^n \to \mathbb{R}^m \\
\frac{df}{dx} \in \mathbb{R}^{m \times n} &= \text{ Jacobian} \\
c^T x &= c_1 x_1 + ... + c_n x_n \\
\frac{dc^T x}{dx} &= \begin{bmatrix} c_1 \\ ... \\ c_n \end{bmatrix} = c \\
\frac{dx^T Q x}{dx} &= \left(Q + Q^T\right)x
\end{aligned}
$$

$$
\begin{aligned}
&\min_x \left\|Ax - b\right\|^2 \\
\left(Ax - b\right)^T\left(Ax - b\right) &= x^T\left(A^T A\right)x - 2b^T Ax + b^T b \\
\frac{d \text{ above}}{dx} &= 2A^T Ax - 2A^T b = 0 \\
A^T Ax - A^T b = 0& \\
x_{opt} &= \left(A^T A\right)^{-1}A^T b
\end{aligned}
$$

4 Lecture 4

4.1 Bayes Classifier

$$(X,Y) \sim \mathbb{P}_{XY}$$
$$\eta(x) := \mathbb{P}\{Y=1|X=x\}$$

$$f^\star(x) = \begin{cases} 1 & \text{if } \eta(x) \geqslant \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{P}\{f^\star(X) \neq Y\} = \mathbb{E}_x\left[\min(\eta(X), 1-\eta(X))\right]$$
$$\{(X_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \mathbb{P}_{XY}$$

4.2 Nearest Neighbor Classifier

new unlabeled example x,

$$i_{1nn(x)} = \arg\min_{i=1\ldots n} \|x - x_i\|$$
$$\hat{y} = y_{i_{1nn(x)}} \leftarrow f_{1nn}(x)$$

Theorem 3. *The following inequality holds,*

$$\lim_{n\to\infty} \mathbb{P}\{f_{1nn}(X) \neq Y\} \leqslant \mathbb{E}\left[2\eta(X)(1-\eta(X))\right] \leqslant 2\mathbb{P}\{f^\star(X) \neq Y\}$$

Theorem 4. *Let* $R_k^\infty = \lim_{n\to\infty} \mathbb{P}\{f_{knn}(X) \neq Y\}$

$$\mathbb{P}\{f^\star(X) \neq Y\} \leqslant R_k^\infty \leqslant R_l^\infty \textit{ for } l < k$$

4.3 KNN Classifier

Theorem 5. *(Stone '77) Let* $k \to \infty$ *and* $\frac{k}{n} \to 0$ *as* $n \to \infty$, *then,*

$$\lim_{n\to\infty} \mathbb{P}\{f_{1nn}(X) \neq Y\} = \mathbb{P}\{f^\star(X) = Y\}$$

$$n \to k = \sqrt{n}$$
$$\mathbb{P}\{f_{1nn}(X) \neq Y\} - \mathbb{P}\{f^\star(X) = Y\} \to 0 \text{ as } n \to \infty$$

4.4 Histogram Classifier

$$X_i \in [0,1]^d, Y_i \in \{0,1\}$$
$$d = 2$$

"bin", $\{B_i\}_{j=1}^{M}$ bins, $M = m^d$

$$\hat{p}_j = \frac{\sum_{i=1}^{n} \mathbb{1}_{x_i \in B_j, y_i = 1}}{\sum_{i=1}^{n} \mathbb{1}_{x_i \in B_j}}, j = 1, ...M$$

if $x \in B_j$ and $\hat{p}_j \geqslant \frac{1}{2}$ then label 1, otherwise 0.

$$\hat{\eta}_n(x) = \sum_{j=1}^{M} \hat{p}_j \mathbb{1}_{x \in B_j}$$

$$\hat{f}_n(x) = \begin{cases} 1 & \text{if } \hat{\eta}_n(x) \geqslant \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Theorem 6. *Let $M \to \infty$ and $\frac{n}{M} \to \infty$ as $n \to \infty$. Then,*

$$\mathbb{P}\left\{\hat{f}_n(X) \neq Y\right\} \to \mathbb{P}\{f^\star(X) \neq Y\}$$

Lemma 1. *,*

$$\mathbb{P}\left\{\hat{f}_n(X) \neq Y\right\} - \mathbb{P}\{f^\star(X) \neq Y\} \leqslant 2\mathbb{E}\left[|\hat{\eta}_n(x) - \eta(x)|\right]$$

$$p_j = \mathbb{P}\{Y = 1 | X \in B_j\}$$
$$p_j = \frac{\int_{B_j} \eta(x) p_x(x) dx}{\int_{B_j} p_x(x) dx}$$

$$\eta^{-x} = \sum_{j}^{M} p_j \mathbb{1}_{x \in B_j}$$

$$\mathbb{E}\left[|\eta(x) - \hat{\eta}_n(x)|\right] \leqslant \underbrace{\mathbb{E}\left[|\eta(x) - \eta^{-n(x)}|\right]}_{\text{Bias },\to 0 as M \to \infty} + \underbrace{\mathbb{E}\left[|\eta^{-x} - \hat{\eta}_n(x)|\right]}_{\text{Variance}}$$

5 Lecture 5

5.1 Bayes Classifier

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$$

$$f^{\star}(x)=\begin{cases}1 & \text{if } \eta(x)\geqslant\frac{1}{2}\\ 0 & \text{otherwise}\end{cases}$$

Ideal Histogram Classifier

$$\begin{aligned}X\in[0,1]^{d}, Y&=\{0,1\}\\ M&=m^{d}\text{ "Bins" },\{B_j\}_{j=1}^{\infty}\\ p_j&=\frac{\mathbb{E}\left[\mathbb{1}_{X\in B_j,Y=1}\right]}{\mathbb{E}\left[\mathbb{1}_{X\in B_j}\right]}\\ &=\frac{\int_{B_j}\eta(x)\,p(x)\,dx}{\int_{B_j}p(x)\,dx}\\ \bar{\eta}(x)&=\sum_{j=1}^{M}p_j\mathbb{1}_{x\in B_j}\end{aligned}$$

$$\bar{f}(x)=\begin{cases}1 & \text{if } \bar{\eta}(x)\geqslant\frac{1}{2}\\ 0 & \text{otherwise}\end{cases}$$

Empiritcal Histogram Classifier

$$\begin{aligned}\hat{p}_j&=\frac{\sum_{i=1}^{n}\mathbb{1}_{x_i\in B_j,y_i=1}}{\sum_{i=1}^{n}\mathbb{1}_{x_i\in B_j}}\\ \{(x_i,y_j)\}&\overset{iid}{\sim}P_{xy}\\ \hat{\eta}(x)&=\sum_{j=1}^{M}\hat{p}_j\mathbb{1}_{x\in B_j}\end{aligned}$$

$$\hat{f}(x)=\begin{cases}1 & \text{if } \hat{\eta}(x)\geqslant\frac{1}{2}\\ 0 & \text{otherwise}\end{cases}$$

A1 $p(x)\geqslant c>0\ \forall\ x$
A2 η is uniformly continuous

$$p(x)\geqslant c>0 \text{ and } \frac{n}{M}\to\infty\Rightarrow N(x)\to^{as}\infty$$

Theorem 7. *If $M\to\infty,\frac{n}{M}\to\infty$ as $n\to\infty$*

$$\underbrace{\mathbb{P}\left\{\hat{f}(X)\neq Y\right\}}_{\mathbb{E}\left[\mathbb{1}_{\hat{f}(X)\neq Y}\right]}\to\underbrace{\mathbb{P}\left\{f^{\star}(X)\neq Y\right\}}_{\mathbb{E}\left[\mathbb{1}_{f^{\star}(X)\neq Y}\right]}$$

Proof. $\mathbb{P}\left\{\hat{f}(X) \neq Y\right\} - \mathbb{P}\{f^{\star}(X) \neq Y\} \leqslant 2\mathbb{E}\left[|\eta(X) - \hat{\eta}(X)|\right]$

$$\begin{aligned}
\mathbb{E}\left[|\eta(X) - \hat{\eta}(X)|\right] &= \mathbb{E}\left[|\eta(X) - \bar{\eta}(X) + \bar{\eta}(X) - \hat{\eta}(X)|\right] \\
&\leqslant \underbrace{\mathbb{E}\left[|\eta(X) - \bar{\eta}(X)|\right]}_{\text{deterministic error}} + \underbrace{\mathbb{E}\left[|\bar{\eta}(X) - \hat{\eta}(X)|\right]}_{\text{stochastic error } \to 0, as n \to \infty}
\end{aligned}$$

where,

$$\begin{aligned}
\mathbb{E}\left[|\eta(X) - \bar{\eta}(X)|\right] &= \int |\eta(x) - \bar{\eta}(x)| p(X)\, dx \\
&= \sum_{j=1}^{M} \int_{B_j} \underbrace{|\eta(x) - \bar{\eta}(x)|}_{\leqslant \varepsilon_m \to 0} p(x)\, dx \\
&\leqslant \varepsilon_m
\end{aligned}$$

and,

$$\mathbb{E}\left[|\bar{\eta}(X) - \hat{\eta}(X)|\right] = \mathbb{E}\left[\mathbb{E}\left[|\bar{\eta}(x) - \frac{K(x)}{N(x)}||X = x\right]\right]$$

where,
x let $B(x)$ be its bin

$$\begin{aligned}
K(x) &= \sum_{i=1}^{n} \mathbb{1}_{x_i \in B(x), y_i = 1} \\
N(x) &= \sum_{i=1}^{n} \mathbb{1}_{x_i \in B(x)}
\end{aligned}$$

$K(x)|N(x) = n_x \sim \text{Binomial}(n_x, \bar{\eta}(x))$

$$\begin{aligned}
\mathbb{E}\left[K(x)|N(x) = n_x\right] &= n_x \bar{\eta}(x) \\
\mathbb{E}\left[\frac{K(x)}{N(x)}|N(x) = n_x\right] &= \frac{n_x \bar{\eta}(x)}{n_x} = \bar{\eta}(x) \\
\mathbb{E}\left[\left(\bar{\eta}(x) - \frac{K(x)}{N(x)}\right)^2 |N(x) = n_x\right] &= \mathbb{E}\left[\frac{1}{(n_x)^2}\left(n_x \bar{\eta}(x) - K(x)\right)^2 |N(x) = n_x\right] \\
&= \frac{1}{(n_x)^2} \mathbb{E}\left[\left(n_x \bar{\eta}(x) - K(x)\right)^2 |X = x\right] \\
&= \frac{1}{(n_x)^2}\left(n_x \bar{\eta}(x)\left(1 - \bar{\eta}(x)\right)\right) \\
&= \frac{\bar{\eta}(x)\left(1 - \bar{\eta}(x)\right)}{n_x}
\end{aligned}$$

Recall, $\mathbb{E}\left[Z^2\right] \geqslant \left(\mathbb{E}[Z]\right)^2$ Jensen's

$$\mathbb{E}\left[|\bar{\eta}(X) - \hat{\eta}(X)||X = x\right] \leqslant \sqrt{\frac{\bar{\eta}(x)\left(1 - \bar{\eta}(x)\right)}{n_x}}$$

$\square$

$$N(x) \propto \frac{n}{M}$$
$$\mathbb{E}\left[|\bar{\eta}(X) - \hat{\eta}(X)|\right] = O\left(\sqrt{\frac{M}{n}}\right)$$

A3: η is 1-Lipschitz

$$|\eta(x) - \eta(x')| \leqslant \|x - x'\| \;\forall\, x, x' \in [0,1]^d$$

if $x, x' \in B_j, \|x - x'\| \leqslant \dfrac{\sqrt{d}}{m} =: \varepsilon_m$

$$\begin{aligned}
\mathbb{P}\left\{\hat{f}(X) \neq Y\right\} - \mathbb{P}\left\{f^\star(X) \neq Y\right\} &= O\left\{\max\left(\sqrt{\frac{m^d}{n}}, \frac{\sqrt{d}}{m}\right)\right\} \\
\sqrt{\frac{m^d}{n}} &= \frac{\sqrt{d}}{m} \\
&\Rightarrow m^{d+2} = dn \\
&\Rightarrow m = (dn)^{\frac{1}{d+2}} \\
\mathbb{P}\left\{\hat{f}(X) \neq Y\right\} - \mathbb{P}\left\{f^\star(X) \neq Y\right\} &= O\left(\frac{\sqrt{d}}{(dn)^{\frac{1}{d+2}}}\right) \\
&= O\left(\frac{1}{n^{\frac{1}{d+2}}}\right) \\
&= O\left(n^{-\frac{1}{d+2}}\right)
\end{aligned}$$

Curse of dim

5.2 Multivariate Normal Distribution

$$\begin{aligned}
X &\in \mathbb{R}^d \\
p(x) &= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right) \\
\mu \in \mathbb{R}^d, \mathbb{E}[X] = \mu, \Sigma &= \mathbb{E}\left[(X - \mu)(X - \mu)^T\right] \in \mathbb{R}^{d \times d} \\
\Sigma_{ij} &= \mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]
\end{aligned}$$

Covariation of X_i, X_j

$$X \sim N(\mu, \Sigma)$$

If $x \sim N(\mu, \Sigma)$, then $Ax + b \sim N\left(A\mu + b, A\Sigma A^T\right)$
Bivariate: $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mathbb{E}[X] = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$\Sigma = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

MVN Classifier
Suppose $X|Y = l \sim \underbrace{N(\mu_l, \Sigma_l)}_{\text{class-conditional distribution of } x}, l = 0, 1, ..., k-1$

$$\mathbb{P}\{Y = l\} = \frac{1}{K}$$
$$\mathbb{P}\{X|Y = l\}\, \mathbb{P}\{Y = l\}$$

Bayes Classifier

$$f^\star(x) = \arg\max_l p(x|y = l)\, \mathbb{P}\{Y = l\}$$
$$K = 2$$

$$f^\star(x) = \begin{cases} 1 & \text{if } \underbrace{\log\left(\frac{p(x|y=1)}{p(x|y=0)}\right)}_{\text{log likelihood ratio}} \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

Log LR

$$\log\left(\frac{\frac{1}{\sqrt{(2\pi)^d |\Sigma_1|}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)\right)}{\frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right)}\right)$$
$$= \underbrace{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) + \frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)}_{\text{quadratic in } x} + \text{const}$$

If $\Sigma_0 = \Sigma_1 = \Sigma$,

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \underbrace{2(\mu_1 - \mu_0)^T \Sigma^{-1}}_{w} x \geqslant \underbrace{\mu_0^T \Sigma \mu_0 - \mu_1^T \Sigma \mu_1}_{t} \\ 0 & \text{otherwise} \end{cases}$$

Linear classifier $w^T x > t$

$$\begin{aligned} &\{x_i, y_i\}_{i=1}^n \\ \hat{\mu}_1 &= \frac{1}{n_1} \sum_{i \in Y_i = 1} x_i, n_1 = \text{ number of } Y_i = 1 \\ \hat{\Sigma}_1 &= \frac{1}{n_1} \sum_{i \in Y_i = 1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T \end{aligned}$$

6 Lecture 6

6.1 Generative Models

$Y = 0$ or $Y = 1$

$$\mathbb{P}\{Y = 0\} = 1 - \mathbb{P}\{Y = 1\}$$

$\mathbb{P}\{X|Y = 0\}, P\{X|Y = 1\}$, Class-conditional distributions

$$\begin{aligned} X|Y = 0 &\sim N(\mu_0, \Sigma_0) \\ X|Y = 1 &\sim N(\mu_1, \Sigma_1) \end{aligned}$$

Region $\hat{y} = 0$ is R_0, region $\hat{y} = 1$ is $R_1, x \in X$

$$\begin{aligned} X &= R_0 \cup R_1 \\ (X, Y) \sim \mathbb{P}_{x,y} &= \mathbb{P}\{X = x, Y = y\} = \mathbb{P}\{X|Y = y\}\mathbb{P}\{Y = y\} \\ \text{Total Err } &= c_{10}\mathbb{P}\left\{Y = 0, \hat{Y} = 1\right\} + c_{01}\mathbb{P}\left\{Y = 1, \hat{Y} = 0\right\}, c_{10}, c_{01} > 0 \\ &= c_{10} \int_{R_1} \underbrace{p(x|y = 0)\, p(y = 0)}_{\geqslant 0} dx + c_{01} \int_{R_0} \underbrace{p(x|y = 1)\, p(y = 1)}_{\geqslant 0} dx \end{aligned}$$

Aside,
X with density $p(x)$, A sub $X, \mathbb{P}\{X \in A\} = \int_A p(x)\, dx$
If $p(x|y = 0)\, p(y = 0) > p(x|y = 1)\, p(y = 1)$
Then put x in R_0
Conversely if $<$, then x in R_1

$$\begin{aligned} \frac{p(x|y = 1)\, p(y = 1)}{p(x|y = 0)\, p(y = 0)} &\gtrless_{\hat{y}=0}^{\hat{y}=1} 1 \\ \text{LR } = \frac{p(x|y = 1)}{p(x|y = 0)} &\gtrless_{\hat{y}=0}^{\hat{y}=1} \frac{c_{10}\, p(y = 0)}{c_{01}\, p(y = 1)} \end{aligned}$$

$$\text{LRT } = \frac{p(x|y=1)}{p(x|y=0)} \gtrless_{\hat{y}=0}^{\hat{y}=1} t$$

6.2 Constrain one type of error

$$\min \mathbb{P}\{\hat{y}=0, Y=1\} \text{ such that } \mathbb{P}\{\hat{y}=1, Y=0\} \leqslant \alpha < 1$$

Neyman-Pearson decision
Solution:

$$\frac{p(x|y=1)}{p(x|y=0)} \gtrless_{\hat{y}=0}^{\hat{y}=1} t_\alpha$$

Example
If $X|Y=j \sim N(\mu_j, \Sigma_j)$

$$\begin{aligned} p(x|y=j) &= \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)} \\ \log(p(x|y=j)) &= -\frac{1}{2} \underbrace{(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}_{\text{Mahanalobis distance}} + \text{constant} \end{aligned}$$

Special case:

$$\Sigma_0 = \Sigma_1 = \Sigma$$

Linear classifier
$\{(X_i, Y_i)\} \overset{iid}{\sim} \mathbb{P}_{X,Y}$ MVN cross conditionals

$$\begin{aligned} j &= 0, 1 \\ \Rightarrow \hat{\mu}_j &= \frac{1}{\#\{i : y_i = j\}} \sum_{i: y_i = j} x_i \\ \hat{\Sigma}_j &= \frac{1}{\#\{i : y_i = j\}} \sum_{i: y_i = j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T \\ \mu_j &= \mathbb{E}[X|Y=j] \\ \Sigma_j &= \mathbb{E}\left[(X-\mu_j)(X-\mu_j)^T |Y=j\right] \end{aligned}$$

Is there a natural notion of "distance" for general class-conditional distributions?

$$\begin{aligned} p_0(x) &= \mathbb{P}\{X|Y=0\} \\ p_1(x) &= \mathbb{P}\{X|Y=1\} \\ \log \text{ LR } &= \log \frac{P_1(x)}{P_0(x)} = \Lambda(x) \gtrless_{\hat{y}=0}^{\hat{y}=1} 0 \end{aligned}$$

if $X \sim q$ (q may be P_0 or P_1 or even something else)

What do we expect log LR to be?

$$\begin{aligned}
\mathbb{E}_{q[\Lambda(X)]} &= \int q(x)\log\left(\frac{p_1(x)}{p_0(x)}\right)dx \\
&= \int q(x)\log\left(\frac{p_1(x)}{p_0(x)}\cdot\frac{q(x)}{q(x)}\right)dx \\
&= \int q(x)\log\left(\frac{q(x)}{p_0(x)}\right)dx - \int q(x)\log\left(\frac{q(x)}{p_1(x)}\right)dx \\
&= D(q\|p_0) - D(q\|p_1)
\end{aligned}$$

KLD of p_i from q, Kullback-Leibler divergence

$$D(q\|p_0) \gtrless^{\hat{y}=1}_{\hat{y}=0} D(q\|p_1)$$

KL for MVN

$$D(N(\mu_0,\Sigma_0)\,\|N(\mu_1,\Sigma_1)) = \frac{1}{2}\left(tr\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - d + \log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)\right)$$

$$D(q\|p) \geqslant 0$$

7 Lecture 7

7.1 Optimal Classification

$$(X,Y)\sim\mathbb{P}_{X,Y}$$

$p(x,y)$ be joint prob density or mass function

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

explicit notation

$$\begin{aligned}
p(y|x) &= \mathbb{P}\{Y=y|X=x\} \\
\eta(x) &= \mathbb{P}\{Y=1|X=x\} \\
1-\eta(x) &= \mathbb{P}\{Y=0|X=x\}
\end{aligned}$$

min prob of error classifier

$$f^\star(x) = \begin{cases} 0 & \text{if } \eta(x)\geqslant\frac{1}{2} \\ 0 & \text{otherwise}\end{cases}$$

$$\eta(x)\geqslant\frac{1}{2} \Leftrightarrow \frac{\eta(x)}{1-\eta(x)}\geqslant 1$$

$$
\begin{aligned}
\frac{\eta(x)}{1-\eta(x)} &= \frac{\mathbb{P}\{Y=1|X=x\}}{\mathbb{P}\{Y=0|X=x\}} = \frac{p(y=1|x)}{p(y=0|x)} \\
&= \frac{\dfrac{p(x,y=1)}{p(x)}}{\dfrac{p(x,y=0)}{p(x)}} = \frac{p(x,y=1)}{p(x,y=0)} \\
&= \frac{p(x|y=1)\,p(y=1)}{p(x|y=0)\,p(y=0)} = \text{ LR}
\end{aligned}
$$

$$
\begin{aligned}
p(x,y) &= p(y|x)\,p(x) \\
&= p(x|y)\,p(y)
\end{aligned}
$$

class-conditional distribution

$$
\begin{aligned}
p_0(x) &= p(x|y=0) \\
p_1(x) &= p(x|y=1) \\
\Lambda(x) &= \log\left(\frac{p_1(x)}{p_0(x)}\right) \\
\Lambda &\gtrless^{\hat{y}=1}_{\hat{y}=0} 0
\end{aligned}
$$

$X \sim q$ prob dist ("test")
$\Lambda(X)$ is a real-valued random variable

$$
\begin{aligned}
\Lambda(X) &= \underbrace{\mathbb{E}[\Lambda(X)]}_{\text{deterministic, a number}} + \underbrace{(\Lambda(X)-\mathbb{E}[\Lambda(X)])}_{\text{zero-mean random variable}} \\
\mathbb{E}[\Lambda(X)] &= \int \Lambda(x)\,q(x)\,dx \\
&= \int q(x)\log\left(\frac{p_1(x)}{p_0(x)}\cdot\frac{q(x)}{q(x)}\right)dx \\
&= \underbrace{\int q(x)\log\left(\frac{q(x)}{p_0(x)}\right)dx}_{D(q\|p_0)} - \underbrace{\int q(x)\log\left(\frac{q(x)}{p_1(x)}\right)dx}_{D(q\|p_1)}
\end{aligned}
$$

Kullback-Leibler (KL) divergences

Lemma 2. *$D(q\|p) \geqslant 0$ for any q,p distributions, $D(q\|q)=0$.*

Proof. ,

$$
\begin{aligned}
D(q\|p) &= \int q(x)\log\left(\frac{q(x)}{p(x)}\right)dx \\
&= -\int q(x)\log\left(\frac{p(x)}{q(x)}\right)dx \\
&= -\mathbb{E}_q\left[\log\left(\frac{p(x)}{q(x)}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\geqslant -\log\left(\mathbb{E}\left[\frac{p(x)}{q(x)}\right]\right)\\
&= -\log\left(\int q(x)\,\frac{p(x)}{q(x)}dx\right)\\
&= -\log(1)\\
&\geqslant 0
\end{aligned}
$$

Jensen's Inequality. If f is convex, then $\mathbb{E}[f(Z)] \geqslant f(\mathbb{E}[Z])$ □

$$
D(q\|p_0) - D(q\|p_1)
$$

Case 1 : $q = p_1$
Case 2 : $q = p_0$

Example 1. $X|Y = 0 \sim N(-\mu, 1), X|Y = 1 \sim N(\mu, 1)$

$$
\begin{aligned}
\Lambda(x) &= \log\left(\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+\mu)^2}{2}}}\right)\\
&= -\frac{(x-\mu)^2}{2} + \frac{(x+\mu)^2}{2}\\
&= \frac{2\mu x + 2\mu x}{2}\\
&= 2\mu x \gtrless^{\hat{y}=1}_{\hat{y}=0}\\
\Lambda(x) &= 2\mu x
\end{aligned}
$$

$$
\begin{aligned}
X &\sim p_1\\
\Lambda(x) = 2\mu X &\sim N\left(2\mu^2, 4\mu^2\right), \sigma = 2\mu
\end{aligned}
$$

μ bigger is better
MVN: $x|y = j \sim N(\mu_j, \Sigma), j = 0, 1$

$$
\begin{aligned}
D(p_0\|p_1) &= D(p_1\|p_0)\\
&= \frac{1}{2}(\mu_1 - \mu_0)^T\,\Sigma^{-1}(\mu_1 - \mu_0)\\
D(p_1\|p_0) &= D(p_0\|p_1)\\
&= 2\mu^2
\end{aligned}
$$

8 Lecture 8

8.1 Bayes Classifier (minimum prob of err)

$$f^\star(x) = \begin{cases} +1 & \text{if } \mathbb{P}(y=1|X=x\} \geqslant \dfrac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

or

$$\begin{cases} +1 & \text{if } \log\left(\dfrac{p(x|y=1)}{p(x|y=-1)}\right) \geqslant \log\left(\dfrac{p(y=-1)}{p(y=+1)}\right) \\ -1 & \text{otherwise} \end{cases}$$

8.2 Nearest Neighbor Classifier

$$\{(X_i, Y_i)\} \overset{iid}{\sim} \mathbb{P}_{XY}$$
$$f_{1nn}(x) = y_{i_x}, i_x = \arg\min_i \|x - x_i\|$$
$$\lim_{n\to\infty} \mathbb{P}\{f_{1nn}(X) \neq Y\} \leqslant 2\mathbb{P}\{f^\star(X) \neq Y\}$$

Example 2. $x \in \{-1,1\}^d, y \in \{-1,1\}, y = x_1$, and $x_2, ..., x_d \overset{iid}{\sim} \pm 1$ with probability $\dfrac{1}{2}$.

$$\text{Bayes Error } = 0$$
$$n = 2, x = (+1, +1, ..., +1)$$
$$d = 2, x = (1, 1)$$

Possible cases for $(x_1, +1)$ and $(x_2, -1)$,

1	-1	correct
-1	1	incorrect
1	1	correct
-1	-1	incorrect
1	-1	tie
-1	1	tie
1	-1	correct
-1	-1	incorrect

$$\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$$
$$\mathbb{P}\{f_{1nn}(X) \neq Y\} \to \frac{1}{2} \text{ as } d \to \infty$$

8.3 Generative Model Plug-in Classifier

$$p(y=+1) = p(y=-1)$$
$$X|Y=+1 \sim N(\theta, I)$$

$$X|Y=-1 \sim N(-\theta, I)$$

Bayes Classifier

$$-\frac{1}{2}\|x-\theta\|^2+\frac{1}{2}\|x+\theta\|^2-\frac{1}{2}\|x-\theta\|^2+\frac{1}{2}\|x+\theta\|^2$$

$$f^\star(x)=\begin{cases} +1 & \text{if } x^T\theta>0 \\ -1 & \text{if } x^T\theta<0 \end{cases}$$

Bayes Err Rate

$$\mathbb{P}\{f^\star(X)\neq Y\}=\frac{1}{2}\mathbb{P}\left\{x^T\theta>0|Y=-1\right\}+\frac{1}{2}\mathbb{P}\left\{x^T\theta<0|Y=+1\right\}$$

$$\mathbb{P}\left\{x^T\theta>0|Y=-1\right\}, x^T\theta \sim N\left(-\|\theta\|^2, \|\theta\|^2\right)$$

$$\begin{aligned}\mathbb{P}\{Z>0\} &= \mathbb{P}\left\{Z'>\|\theta\|^2\right\}, Z\sim N\left(-\|\theta\|^2,\|\theta\|^2\right), Z'\sim N\left(0,\|\theta\|^2\right)\\ &\leqslant \frac{\|\theta\|^2}{(\|\theta\|)^4}\\ &= \frac{1}{\|\theta\|^2}\end{aligned}$$

Use Markov,

$$\mathbb{P}\{Z>t\}\leqslant\mathbb{P}\left\{Z^2>t^2\right\}\leqslant\frac{\mathbb{E}\left[Z^2\right]}{t^2}$$

Plug-in,

$$\hat\theta=\frac{1}{n}\sum_{i=1}^n y_i x_i$$

$$\hat f(x)=\begin{cases} +1 & \text{if } x^T\hat\theta>0 \\ -1 & \text{if } x^T\hat\theta<0 \end{cases}$$

What is the distribution of $x^T\hat\theta|Y=-1$

$$\begin{aligned} X &= -\theta+e_1, e_1\sim N(0,I)\\ \hat\theta &= \theta+e_2, e_2\sim N\left(0,\frac{1}{n}I\right)\end{aligned}$$

Ignoring constant factors,

$$\mathbb{P}\left\{x^T\hat\theta>0|Y=-1\right\}\leqslant\frac{1}{\|\theta\|^2}+\frac{d^2}{n}\frac{1}{\|\theta\|^4}$$

$$\frac{d^2}{n}\frac{1}{\|\theta\|^4}\approx\frac{1}{\|\theta\|^2}$$

$$n\approx\left(\frac{\|\theta\|}{d}\right)^2$$

8.4 Maximum Likelihood Estimation

$$x_1, ..., x_n \sim q$$
$$q \in \{p_\theta\}_{\theta \in \Theta}$$

Example 3. $x_1, ..., x_n \overset{iid}{\sim} N(\theta, I)$ for some $\theta \in \mathbb{R}^d$

Two approaches

1. Method of Moments

$$\begin{aligned} \mu_f &= \int f(x)\, q(x)\, dx, \text{ for any } f \\ \hat{\mu}_f &= \frac{1}{n} \sum_{i=1}^{n} f(x_i) \\ \mu_f(\theta) &= \int f(x)\, p_\theta(x)\, dx \end{aligned}$$

find θ that minimizes,

$$|\hat{\mu}_f - \mu_f(\theta)| \to \hat{\theta}$$

for one or more functions f.

2. MLE

$$\begin{aligned} \hat{\theta} &= \arg\max_\theta \prod_{i=1}^{n} p_\theta(x_i) \\ &= \arg\min_\theta - \sum_{i=1}^{n} \log p_\theta(x_i) \end{aligned}$$

9 Lecture 9

9.1 Maximum Likehood

$$x_1, ..., x_n \overset{iid}{\sim} q$$

Models $\{P_\theta\}_{\theta \in \Theta}$

$$L(\theta) = \log\left(\prod_{i=1}^{n} P_\theta(x)\right)$$

when viewed as function of θ is called the likehood of θ.

$$\hat{\theta} = \arg\max_\theta L(\theta)$$

Example 4. $x|\theta \sim N(\theta, I), \theta \in \mathbb{R}^d$

$$\begin{aligned}
\log p(x|\theta) &= \log\left(\frac{1}{\sqrt{2\pi d}}\exp\left(-\frac{1}{2}(x_1-\theta)^T(x_1-\theta)\right)\right)\\
&= -\frac{1}{2}(x_1-\theta)^T(x_1-\theta) + \text{ const}\\
&= -\frac{1}{2}\left(x_1^T x_1 - 2x_1^T\theta + \theta^T\theta\right)\\
\hat{\theta} &= \arg\max_{\theta\in\mathbb{R}^d} -\sum_{i=1}^{n}\frac{1}{2}(x_1-\theta)^T(x_1-\theta)\\
&\Rightarrow \sum_{i=1}^{n}(x_i-\theta) = 0\\
&\Rightarrow \sum x_i = \sum \theta\\
&\Rightarrow \sum x_i = n\theta\\
&\Rightarrow \hat{\theta} = \frac{1}{n}\sum x_i
\end{aligned}$$

Example 5. $x|\theta \sim$ Poiss $(\theta), \theta > 0$

$$\begin{aligned}
p(x|\theta) &= e^{-\theta}\frac{\theta^x}{x!}, x = 1, 2, ...\\
\log p(x|\theta) &= -\theta + x\log\theta - \log x!\\
&x_1, x_2, ..., x_n\\
&\max_\theta \sum_i(-\theta + x_i\log\theta)\\
&\Rightarrow \sum_i(-\theta + x_i) = 0\\
&\Rightarrow \hat{\theta} = \frac{1}{n}\sum x_i
\end{aligned}$$

MLE $\Rightarrow \hat{\theta}, P_{\hat{\theta}}$

Questions:

1. In what sense is $P_{\hat{\theta}} a$ good model for q
2. If $q = P_{\theta^\star}$, then we hope that $\hat{\theta} \to \theta^\star$ as $n \to \infty$

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta \sum_{i=1}^{n}\log p(x_1|\theta)\\
&= \arg\min_\theta -\sum_{i=1}^{n}\log p(x_i|\theta) + \log q(x_i)\\
&= \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(x_i)}{p(x_i|\theta)}\\
\frac{1}{n}\sum_{i=1}^{n}\log\frac{q(x_i)}{p(x_i|\theta)} &\to_{n\to\infty} D(q\|P_\theta) \text{ a.s.}\\
\theta^\star &= \arg\min_\theta D(q\|p_\theta)
\end{aligned}$$

$$
\begin{aligned}
0 \geqslant \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(x_i)}{p\left(x_i|\hat{\theta}\right)} - \sum_{i=1}^{n}\log\frac{q(x_i)}{p(x_i|\theta^\star)} \\
= D\left(q|p_{\hat{\theta}}\right) - D\left(q\|p_{\theta^\star}\right) \\
D\left(q\|p_{\theta^\star}\right) \leqslant D\left(q\|p_{\hat{\theta}}\right) \leqslant (\approx) D\left(q\|p_{\theta^\star}\right)
\end{aligned}
$$

Claim: $D\left(q\|p_{\hat{\theta}}\right) \to D\left(q\|p_{\theta^\star}\right)$ as $n \to \infty$

9.2 Central Limit Theorem

If $z_1, .., z_n$ are zero mean, id with variance σ^2, then

$$
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right) \overset{\text{assmp dist}}{\sim} N\left(0, \sigma^2\right)
$$

Theorem 8. *$x_1, x_2, ..., x_n \overset{iid}{\sim} p_{\theta^\star}$. Let $L(\theta) = \sum_{i=1}^{n}\log p(x_i|\theta)$. Assume $\frac{\partial L}{\partial \theta_j}$ and $\frac{\partial L}{\partial \theta_j \partial \theta_k}$ exist $\forall\ j, k$. Then,*

$$
\sqrt{n}\left(\hat{\theta}_n - \theta^\star\right) \overset{assmp\ dist}{\sim} N\left(0, I^{-1}(\theta^\star)\right)
$$

where, Fisher Information matrix,

$$
[I(\theta^\star)]_{j,k} = -\mathbb{E}\left[\left.\frac{\partial \log p(x|\theta)}{\partial \theta_j \partial \theta_k}\right|_{\theta=\theta^\star}\right]
$$

Low curvature $\Rightarrow$ high variance
High curvature $\Rightarrow$ low variance in MLE

Proof. 1-dim case

$$
L(\theta) = \sum_{i=n}\log p(x_i|\theta)
$$

Taylor's Series (Mean Value Theorem):

$$
f(x) = f(x_0) + f(x_0) + f'(\bar{x})(x - x_0)
$$

$\bar{x}$ is between x and x_0

$$
\begin{aligned}
0 = L'\left(\hat{\theta}\right) = L'(\theta^\star) + L''\left(\bar{\theta}\right)\left(\hat{\theta} - \theta^\star\right) \\
\Rightarrow \left(\hat{\theta} - \theta^\star\right) = -\frac{L'(\theta^\star)}{L''\left(\bar{\theta}\right)}, \bar{\theta} \text{ is between } \hat{\theta} \text{ and } \theta^\star \\
\Rightarrow \sqrt{n}\left(\hat{\theta} - \theta^\star\right) = -\frac{\frac{1}{\sqrt{n}}L'(\theta^\star)}{\frac{1}{n}L''\left(\bar{\theta}\right)} \\
\frac{1}{\sqrt{n}}L'(\theta^\star) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left.\frac{\partial \log p(x_i|\theta)}{\partial \theta}\right|_{\theta=\theta^\star}
\end{aligned}
$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\left.\frac{\partial \log p\left(x_i|\theta\right)}{\partial \theta}\right|_{\theta=\theta^\star}\right] &= \mathbb{E}\left[\frac{1}{p\left(x_i|\theta\right)}\left.\frac{\partial p\left(x_i|\theta\right)}{\partial \theta}\right|_{\theta=\theta^\star}\right] \\
&= \int \frac{1}{p\left(x_i|\theta\right)}\left.\frac{\partial p\left(x_i|\theta\right)}{\partial \theta}\right|_{\theta=\theta^\star} p\left(x_i|\theta^\star\right) dx_i \\
&= \int 1 \left.\frac{\partial p\left(x_i|\theta\right)}{\partial \theta}\right|_{\theta=\theta^\star} dx_i \\
&= \left.\frac{\partial}{\partial \theta}\left(\int p\left(x_i|\theta\right) dx\right)\right|_{\theta=\theta^\star} \\
&= \left.\frac{\partial 1}{\partial \theta}\right|_{\theta=\theta^\star} \\
&= 0
\end{aligned}$$

and,

$$\begin{aligned}
\frac{\partial^2 \log p\left(x|\theta\right)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left(\frac{1}{p\left(x|\theta\right)}\frac{\partial p\left(x|\theta\right)}{\partial \theta}\right) \\
&= -\frac{1}{p^2\left(x|\theta\right)}\left(\frac{\partial p\left(x|\theta\right)}{\partial \theta}\right)^2 + \frac{1}{p\left(x|\theta\right)}\frac{\partial^2 p\left(x|\theta\right)}{\partial \theta^2}
\end{aligned}$$

In expectations given $\theta = \theta^\star$,

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 \log p\left(x|\theta\right)}{\partial \theta^2}\right] &= \mathbb{E}\left[-\frac{1}{p^2\left(x|\theta\right)}\left(\frac{\partial p\left(x|\theta\right)}{\partial \theta}\right)^2\right] + 0 \\
&= \int -\frac{1}{p^2\left(x|\theta\right)}\left.\left(\frac{\partial p\left(x|\theta\right)}{\partial \theta}\right)^2\right|_{\theta=\theta^\star} p\left(x|\theta^\star\right) dx \\
&= -\mathbb{E}\left[\left.\left(\frac{\partial \log p\left(x|\theta\right)}{\partial \theta}\right)^2\right|_{\theta=\theta^\star}\right] \\
&= -\mathbb{V}\left[\left.\frac{\partial \log p\left(x|\theta\right)}{\partial \theta}\right|_{\theta=\theta^\star}\right]
\end{aligned}$$

□

10 Lecture 10

10.1 Performance of MLE

$$\begin{aligned}
x_1, x_2, ..., x_n &\overset{iid}{\sim} P_{\theta^\star} \\
\hat{\theta}_n &= \arg\min_{\theta} - \sum_{i=1}^{n} \underbrace{\log p\left(x_i|\theta\right)}_{l(\theta)}
\end{aligned}$$

10.2 Fischer Info Matrix

$$I(\theta^\star) = \left\{ \mathbb{E}\left[-\frac{\partial \log p(x|\theta)}{\partial\theta_j \partial\theta_k} \bigg|_{\theta=\theta^\star} \right] \right\}_{j,k}$$

$$\hat{\theta}_n \overset{asymp}{\sim} N\left(\theta^\star, \quad \underbrace{\frac{1}{n} I^{-1}(\theta^\star)}_{\text{error covariance}} \right)$$

Assumptions:

1. $L'(\theta), L''(\theta)$ exist,
2. $\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right] = 0 \;\forall\; \theta$

Example 6. $x_1, ..., x_n \overset{iid}{\sim} N(\theta, \sigma^2), \sigma^2$ known, θ unknown, $\theta \in \mathbb{R}$

$$\hat{\theta} = \frac{1}{n}\sum x_i \sim N\left(\theta^\star, \frac{\sigma^2}{n}\right)$$

$$\frac{\partial \log p(x|\theta)}{\partial\theta} = \frac{2}{\sigma}\left(-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2} + \text{ const} \right) = \frac{x-\theta}{\sigma^2}$$

$$\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2} = -\frac{1}{\sigma^2} \Rightarrow I(\theta^\star) = \frac{1}{\sigma^2}$$

Example 7. $x_i \overset{iid}{\sim}$ Unif $[0, \theta]$

$$\hat{\theta}_n = \max_{i=1,...,n} x_i$$

extremal stats

$$x_i \overset{iid}{\sim} N(0,1)$$

$$\max_{i=1,...,n} x_i \overset{asymp}{\sim} \sqrt{2\log n}$$

$$x_{i^\star} \sim N(\mu, 1), \mu > 0$$

10.3 Sufficient Statistics

$$X \sim \text{ Bernoulli } (\theta)$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}, x \in \{0,1\}$$

$$x_1, ..., x_n \overset{iid}{\sim} \text{ Be } (\theta)$$

$$S = \sum x_i \sim \text{ Bi } (\theta, n)$$

$$p(s = k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$p\left(x_1,...,x_n,S|\theta\right)=\begin{cases} p\left(x_1,...,x_n|\theta\right) & \text{if } \sum x_i=S \\ 0 & \text{otherwise}\end{cases}$$

$$\begin{aligned}
p\left(x_1,...,x_n|s=k,\theta\right) &= \frac{p\left(x_1,...,x_n,s=k|\theta\right)}{p\left(s|\theta\right)} \\
&= \frac{\prod_{i=1}^{n}\theta^{x_i}\left(1-\theta\right)^{1-x_i}}{\binom{n}{k}\theta^k\left(1-\theta\right)^{n-k}} \\
&= \frac{\theta^k\left(1-\theta\right)^{n-k}}{\binom{n}{k}\theta^k\left(1-\theta\right)^{n-k}} \\
&= \frac{1}{\binom{n}{k}}
\end{aligned}$$

$$\begin{aligned}
&\max_\theta p\left(x_1,...,x_n,s|\theta\right) \\
&= \max_\theta p\left(x_1,...,x_n|s\right)p\left(s|\theta\right) \\
&= \max_\theta p\left(\underbrace{s}_{\text{sufficient statistics for }\theta}|\theta\right)
\end{aligned}$$

10.4 Fischer-Neyman Factorization Thm

$$x_1,...,x_n\overset{iid}{\sim}P_\theta$$

$t\left(x_1,...,x_n\right)$ is a sufficient statistic for θ iff

$$p\left(x_1,...,x_n|\theta\right)=\underbrace{a\left(x_1,...,x_n\right)}_{\text{no }\theta\text{ dependence}}b\left(t,\theta\right)$$

MVN $x_i\overset{iid}{\sim}N\left(\mu,\Sigma\right)$, both unknowns, $\theta=\left(\mu,\Sigma\right),x_i\in\text{real}^d,i=1,...,n,nd$
Sufficient Statistics:

$$\begin{aligned}
n\hat{\mu} &= \sum_i x_i \\
n\hat{\Sigma} &= \sum_i\left(x_i-\hat{\mu}\right)\left(x_i-\hat{\mu}\right)^T=t\left(x_1,...,x_n\right) \\
&\approx d+d^2\text{ numbers}
\end{aligned}$$

10.5 Rao-Blackwell Theorem

$x_1, ..., x_n \overset{iid}{\sim} P_{\theta^\star}, t = t(x_1, ..., x_n)$, sufficient for θ,
Let $\underbrace{f(x_1, ..., x_n)}_{\hat{\theta}}$ be an estimator of θ

$$\underbrace{g([t(x_1, ..., x_n)])}_{\text{just depends on } x's \text{ through sufficient statistics}} = \mathbb{E}[f(x_1, ..., x_n) | t(x_1, ..., x_n) = t]$$

$$\mathbb{E}[g(t(x_1, ..., x_n))] = \mathbb{E}[\mathbb{E}[f(x_1, ..., x_n) | t(x_1, ..., x_n) = t]] = \mathbb{E}[f(x_1, ..., x_n)]$$

Then,

$$\mathbb{E}\left[g(t(x_1, ..., x_n))^2\right] \leqslant \mathbb{E}\left[(f(x_1, ..., x_n) - \theta)^2\right]$$

Example 8. $n > 2, x_1, ..., x_n \overset{iid}{\sim} N(\theta, 1), t_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$f(x_1, ..., x_n) = \frac{x_1 + x_2}{2}$$

10.6 Linear Models

predict y from $x \in \mathbb{R}^d$

$$\hat{y} = f\left(x^T w\right)$$

f non-linear, $w =$ weights $\in \mathbb{R}^d$, "weighted sum of features"

$$\hat{y} = \sum_{i=1}^{k} w_j x_j$$

10.7 Linear Regression

$\{(x_i, y_i)\}_{i=1}^{n}$ training data

$$\hat{y}_i = x_i^T w$$

prediction errors

$$\begin{aligned} e_1 &= y_1 - x_1^T w \\ e_2 &= y_2 - x_2^T w \\ &... \\ e_n &= y_n - x_n^T w \end{aligned}$$

Least squares

$$y = \begin{bmatrix} y_1 \\ ... \\ y_n \end{bmatrix}$$

$$x = \begin{bmatrix} x_1^T \\ ... \\ x_n^T \end{bmatrix}$$

$$w = \begin{bmatrix} w_1 \\ ... \\ w_n \end{bmatrix}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

$$\min_w \|y - Xw\|^2$$

$$\begin{aligned} \|y - Xw\|^2 &= (y - Xw)^T (y - Xw) \\ &= y^T y - 2y^T Xw + x^T X^T Xw \\ 0 &= \frac{\partial \text{ above}}{\partial w} = 2X^T y + 2X^T Xw \\ X^T y &= X^T Xw \end{aligned}$$

normal equations
$\hat{w} = \left(X^T X\right)^{-1} X^T y$, assuming invertable

11 Lecture 11

11.1 Linear Models

$$\begin{aligned} e_1 &= y_1 - x_1^T w \\ e_2 &= y_2 - x_2^T w \end{aligned}$$

11.2 Least Squares (LS)

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

$$\left(x^T x\right)^{-1} x^T y = \arg\min_w \|y - xw\|_2^2$$

$$\sum_i \left(y_i - x_i^T w\right)^2 = \sum_i \left(y_i^2 - 2w^T x_i y_i + w^T x_i x_i^T w\right)$$

$$w^T \left(\sum \begin{bmatrix} x_{i1} \\ ... \\ x_{id} \end{bmatrix} y_i\right) \to w^T \begin{bmatrix} \sum x_{i1} y_i \\ ... \\ \sum x_{id} y_i \end{bmatrix}$$

If for example,

$$\begin{aligned}
\sum x_{i1} y_i \text{ large } + &\Rightarrow w_1 > 0 \\
\sum x_{i2} y_i \text{ large } - &\Rightarrow w_2 < 0 \\
\sum x_{i3} y_i = 0 &\Rightarrow w_3 \approx 0
\end{aligned}$$

11.3 MLE Perspective

$$\begin{aligned}
y_i &= x_i^T w + v_i, v_i \overset{iid}{\sim} N(0,1) \\
y_i - x_i^T w &= v_i \sim N(0,1) \\
p(v_i) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} v_i^2} \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(y_i - x_i^T w\right)^2} \\
\log \prod_{i=1}^{n} p(v_i) &= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(y_i - x_i^T w\right)^2} \right) \\
&= -\frac{1}{2} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \text{ const.}
\end{aligned}$$

$p(v_i) = \frac{1}{2} e^{-|v_i|}$ heavy tails

$$\begin{aligned}
&\log \prod_{i=1}^{n} \frac{1}{2} e^{-|y_i - x_i^T w|} \\
&= -\frac{1}{2} \sum_{i=1}^{n} |y_i - x_i^T w| + \text{ const} \\
&\Rightarrow \text{ MLE } \max_w - \sum |y_i - x_i^T w| \\
&\Rightarrow \min_w \sum |y_i - x_i^T w| \\
&\Rightarrow \min_w \|y - x^T w\|_1
\end{aligned}$$

$$z_1, z_2, ..., z_n$$

$\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (z_i - \theta)^2 \Rightarrow \theta$ is mean

$\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} |z_i - \theta| \Rightarrow \theta$ is median

11.4 Binary Classification and Linear Prediction

$$y \in \{0,1\}, \hat{y}_i = x_i^T w$$

$$\mathbb{P}\left\{y_i = 1\right\} = f\left(x_i^T w\right), f : \mathbb{R} \to [0,1]$$

$$\hat{y}_i = \begin{cases} 1 & \text{if } \operatorname{sign}\left(x_i^T w\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

11.5 Bernoulli Distribution

$$\begin{aligned} p\left(y_i\right) &= p_i^{y_i}\left(1 - p_i\right)^{1-y_i} \\ &= f\left(x_i^T w\right)^{y_i}\left(1 - f\left(x_i^T w\right)\right)^{1-y_i} \\ &= \exp\left(y_i \log p_i + \left(1 - y_i\right)\log\left(1 - p_i\right)\right) \\ &= \exp\left(y_i \underbrace{\log\left(\frac{p_i}{1-p_i}\right)}_{\theta_i} + \log\left(1 - p_i\right)\right) \end{aligned}$$

Canonical $\theta_i = \log\left(\dfrac{p_i}{1-p_i}\right)$

$$\begin{aligned} \theta_i &= x_i^T w \\ y_i\theta_i &= y_i x_i^T w \\ &= w x_i y_i^T \\ e^{\theta_i} &= \frac{p_i}{1-p_i} \\ e^{\theta_i}\left(1 - p_i\right) &= p_i \\ e^{\theta_i} &= \left(1 + e^{\theta_i}\right) p_i \\ p_i &= \frac{e^{\theta_i}}{1 + e^{\theta_i}} = \frac{1}{1 + e^{-\theta_i}} \\ f\left(x_i^T w\right) &= \frac{1}{1 + e^{-\theta_i}} \\ &\max_{w \in \mathbb{R}^d} \prod_{i=1}^{n} \exp\left(y_i x_i^T w + \log\left(1 - \frac{1}{1 + e^{-x_i^T w}}\right)\right) \end{aligned}$$

11.6 Logistic Regression

$$\begin{aligned} \theta_i &= x_i^T w \\ &p_i^{y_i}\left(1 - p_i\right)^{1-y_i} \\ p\left(y_i|\theta_i\right) &= \exp\left(y_i \log p_i + \left(1 - y_i\right)\log\left(1 - p_i\right)\right) \\ &= \exp\left(y_i \log\left(\frac{1}{1 + e^{-\theta_i}}\right) + \left(1 - y_i\right)\log\left(\frac{1}{1 + e^{\theta_i}}\right)\right) \end{aligned}$$

$$\log\left(p\left(y_i|\theta_i\right)\right) = \begin{cases} \log\left(\dfrac{1}{1 + e^{-\theta_i}}\right) & \text{if } y_i = 1 \\ \log\left(\dfrac{1}{1 + e^{\theta_i}}\right) & \text{if } y_i = 0 \end{cases}$$

$$z_i = 2y_i - 1 \in \{-1, 1\}$$
$$\sum_{i=1}^{n} \log p\left(z_i|\theta_i\right) = \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{-\theta_i z_i}}\right)$$

MLE, $\theta_i = w^T x_i$

$$\max_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{-w^T x_i z_i}}\right)$$
$$= \min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \underbrace{\log\left(1+e^{-w^T x_i z_i}\right)}_{\text{logistic loss function}}$$

LS,

$$\min_{w} \sum_{i=1}^{n} \underbrace{\left(y_i - x_i^T w\right)^2}_{\text{squared err loss function}}, z_i = 2y_i - 1$$

11.7 Multiclass Problems

$$y_i \in \{1, 2, ..., k\}$$

k weight vectors, $w_1, ..., w_k$

$$l \in \{1, ..., k\}$$
$$p\left(y_i = l\right) = \frac{e^{w_l^T x_i}}{\sum_{j=1}^{k} e^{w_j^T x_i}} \in [0, 1]$$
$$\max_{w_1,...,w_k\in\mathbb{R}^d} \sum_{i=1}^{n} \log\left(\frac{e^{w_l^T x_i}}{\sum_{j=1}^{k} e^{w_j^T x_i}}\right)$$

Softmax.
Multinomial Logistic Regression

12 Lecture 12

12.1 GLMs

Models for $p\left(y|x\right)$ then depend on x only through a linear map $w^T x$. The weight $w \in \mathbb{R}^d$ can be fit by maximum likelihood.

$$p\left(y|x\right) = p\left(y|w^T x\right)$$

Gaussian: $y_i \in \mathbb{R}$,

$$p\left(y_i | w^T x_i\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(y_i - w^T x_i\right)^2\right)$$

Binomial: $y_i \in \{0, 1\}$

$$p\left(y_i | w^T x_i\right) = \exp\left(y_i \log\left(\frac{1}{1+e^{-w^T x_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{w^T x_i}}\right)\right)$$

$$y_i \in \{-1, +1\}$$

$$p\left(y_i | w^T x_i\right) = \exp\left(\log\left(\frac{1}{1+e^{-y_i x_i^T w}}\right)\right)$$

12.2 MLEs

Guassian:

$$\min_{w \in \mathbb{R}^d} - \sum_{i=1}^{n} \underbrace{\frac{1}{2}\left(y_i - w^T x_i\right)^2}_{\text{squared error loss}}$$

$$= \min_w \|y - Xw\|^2$$

If X is full rank,

$$\hat{w} = \left(X^T X\right)^{-1} X^T y$$

Bernoulli:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \underbrace{\log\left(1 + \exp\left(-y_i x_i^T w_i\right)\right)}_{\text{logistic loss}}$$

12.3 Binary Classification

$$y_i \in \{-1, +1\}$$

sq err loss:

$$\left(y_i - w^T x_i\right)^2 = \left(1 - y_i x_i^T w\right)^2$$

logistic loss:

$$\log\left(1 + e^{-y_i x_i^T w}\right)$$

hinge loss:

$$\max\left\{0, 1 - y_i x_i^T w\right\} = \left(1 - y_i x_i^T w\right)^+$$

$0-1$ loss:

$$\mathbb{1}_{y_i x_i^T w < 0}$$

Then,

$$\begin{aligned}\hat{y}_i &= \text{sign}\left(x_i^T w\right)\\ \{(x_i, y_i)\}_{i=1}^n &\overset{iid}{\sim} \mathbb{P}_{xy}\\ \mathbb{P}_{xy}&\left\{y \neq \text{sign}\left(\hat{w}^T x\right)\right\}\end{aligned}$$

12.4 Convex Losses

loss l is convex in w if for any w_0, w_1, and $\lambda \in [0,1]$

$$l\left(\lambda y_i x_i^T w_0 + (1-\lambda) y_i x_i^T w_1\right) \leqslant \lambda l\left(y_i x_i^T w_0\right) + (1-\lambda) l\left(\lambda y_i x_i^T w_1\right)$$

12.5 Least Square

$$\begin{aligned}&\min_w \quad \underbrace{\|y - Xw\|^2}_{(y-Xw)^T(y-Xw)}\\ \frac{\partial \text{ above}}{\partial w} &= -2X^T y + 2X^T Xw\\ &= -2X^T(y - Xw)\end{aligned}$$

Set it to 0,

$$\begin{aligned}X^T y &= X^T X w\\ w^\star &= \left(X^T X\right)^{-1} X^T y\end{aligned}$$

w_1 initial guess (often 0 or random)

$$\begin{aligned}w_{t+1} &= w_t - \gamma \nabla f(w_t), t = 1, 2, ...\\ &= w_t + \underbrace{\gamma}_{\text{step size}} X^T\left(y - X^T w_t\right)\end{aligned}$$

13 Lecture 13

13.1 Gradient Descent and LS

$$\begin{aligned}\min_w \|y - Xw\|^2 &= \min_w \sum_{i=1}^n \left(y_i - x_i^T w\right)^2\\ f(w) &= \|y - Xw\|^2\\ \nabla f(w) &= \frac{\partial}{\partial w}\left((y - Xw)^T (y - Xw)\right)\end{aligned}$$

$$
\begin{aligned}
&= -X^T y + X^T X w \\
&= -X^T (y - Xw) \\
w_{t+1} &= w_t - \gamma \nabla f(w_t), t = 1, 2, ..., \gamma > 0 \\
&= w_t + \gamma X^T (y - X w_t) \\
&= w_t + \gamma \left(X^T - X^T X w_t \right) \\
&= w_t + \gamma X^T X \left(\left(X^T X \right)^{-1} X^T y - w_t \right) \\
&= w_t - \gamma X^T X (w_t - w^\star)
\end{aligned}
$$

Goal:

$$
\begin{aligned}
w_t - w^\star &\to 0 \text{ as } t \to \infty \\
w_{t+1} - w^\star &= w_t - w^\star - \gamma X^T X (w_t - w^\star) \\
&= \left(I - \gamma X^T X \right) \underbrace{(w_t - w^\star)}_{v_t} \\
v_{t+1} &= \left(I - \gamma X^T X \right) v_t \\
&= \left(I - \gamma X^T X \right) \left(I - \gamma X^T X \right) v_{t-1} \\
&= \left(I - \gamma X^T X \right)^t v_1
\end{aligned}
$$

Eigendecomposition of $\left(I - \gamma X^T X \right)$

$$
I - \gamma X^T X = U D U^T
$$

U: orthogonal: $U^T U = I$
D: diagonal $\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}$

$$
\begin{aligned}
D^2 &= \begin{bmatrix} \gamma_1^2 & 0 \\ 0 & \gamma_2^2 \end{bmatrix} \\
\left(U D D^T \right)^2 &= \left(U D U^T \right) \left(U D U^T \right) \\
&= U D I D U^T \\
&= U D^2 U^T
\end{aligned}
$$

Convergence Condition:

$$
\begin{aligned}
|\lambda_i| &< 1 \text{ for } i = 1, 2, ..., d \\
D &= U^T U D U^T U \\
&= U^T \left(I - Y X^T X \right) U \\
&= \left(I - \gamma U^T X^T X U \right) \leftarrow \text{ is diagonal}
\end{aligned}
$$

$U^T X^T X U = \mathrm{diag}\big(\lambda_1\left(X^T X\right) > 0, ..., \lambda_d\left(X^T X\right) > 0\big)$
is positive definite.

$$|\lambda_i| = |1 - \gamma\lambda_i\left(X^T X\right)| < 1$$
$$\gamma\lambda_i\left(X^T X\right) \leqslant 2, i = 1, ..., d$$
$$\Rightarrow \gamma < \frac{2}{\lambda\text{max}\left(X^T X\right)}$$

LS GD:

$$w_{t+1} = w_t - \gamma\frac{\partial}{\partial w}\left(\frac{1}{2}\sum_{i=1}^{n}\left(y_i - x_i^T w\right)^2\right)\Bigg|_{w=w_t}$$

Stochastic GD:

$$w_{t+1} = w_t - \gamma\frac{\partial}{\partial w}\frac{1}{2}\left(y_{i_t} - x_{i_t}^T w\right)^2\Bigg|_{w=w_t}, i_t \sim \text{ Unif } (1, ..., n)$$

Then,

$$\mathbb{E}\left[\frac{\partial}{\partial w}\frac{1}{2}\left(y_{i_t} - X_{i_t}^T w\right)^2\right]$$
$$= \mathbb{E}\left[\frac{\partial}{\partial w}\sum_{i=1}^{n}\left(y_i - X_i^T w\right)\mathbb{1}_{\{i=i_t\}}\right]$$
$$= \frac{\partial}{\partial w}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - X_i^T w\right)^2$$

13.2 SGD with Convex Loss

$$w^\star = \arg\max_{w\in\mathbb{R}^d} = \frac{1}{T}\sum_{i=1}^{T} f_t(w), \text{ convex in } w$$

LS:

$$f_t(w) = \left(y_{i_t} - X_{i_t}^T w\right)^2$$

Logistic:

$$f_t(w) = \log\left(1 + \exp\left(-y_{i_t} x_{i_t}^T w\right)\right)$$

$T \to \infty$, LS:

$$\frac{1}{T}\sum_{i=1}^{T}\Bigg(\underbrace{y_{i_t} - X_{i_t}^T w}_{Z_t \text{ iid}}\Bigg)^2 \to^{as} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T w\right)^2$$
$$\mathbb{E}\left[Z_t\right] = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T w\right)^2$$

Convex iff f is above tangent plane
for any $w, u \in \mathbb{R}^d$

$$f(u) \geqslant f(w) + \nabla^T f(w)(u - w)$$

Example,

$$\begin{aligned} w^2 &= f(w) \\ \frac{\partial}{\partial x} w^2 &= 2w = 0 \\ w &= 0, f(w) = 0, \nabla f(0) = 0 \\ f(u) &\geqslant 0 + 0(u - w) \\ &\geqslant 0 \\ w &= 1, f(1) = 1, \nabla f(1) = 2 \\ f(u) &\geqslant 1 + 2(u - w) \end{aligned}$$

14 Lecture 14

14.1 Supervised ML

$\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \mathbb{P}_{XY}$, unknown
learn to predict y from x

What would we do if we knew $\mathbb{P}_{XY}$
in Binary Classification

$$y \in \{0, 1\} \text{ or } \{-1, +1\}$$

$$\frac{\mathbb{P}\{Y = 1 | X = x\}}{\mathbb{P}\{Y = -1 | X = x\}} \gtrless_{\text{class } -1}^{\text{class } 1} 1$$

14.2 Bayes Rule

$$\begin{aligned} p(y|x) = & \frac{p(x|y)\, p(y)}{p(x)} \\ & \frac{p(x|y = +1)\, p(y = +1)}{p(x|y = -1)\, p(y = -1)} \gtrless_{\text{class } -1}^{\text{class } 1} 1 \\ & \frac{p(x|y = +1)}{p(x|y = -1)} \gtrless_{\text{class } -1}^{\text{class } 1} \frac{p(y = -1)}{p(y = +1)} \end{aligned}$$

$$\begin{aligned} \eta(x) &= p(y = +1|x) \\ \hat{\eta}(x) &\gtrless_{\text{class } -1}^{\text{class } 1} \frac{1}{2} \end{aligned}$$

"plug-in" approach

14.3 Empirical Risk Minization

$\{(x_i, y_i)\}_{i=1}^n$, set of classifiers $\mathcal{F}$

$$f(x) = \pm 1$$
$$\min_{f \in \mathcal{F}} \sum_{i=1}^n \text{ loss } (f(x_i), y_i)$$

$0 - 1$ loss: $\mathbb{1}_{\{f(x_i) \neq y_i\}} = \mathbb{1}_{\{y_i f(x_i) < 0\}}$
sq err loss: $(y_i - f(x_i))^2 = (1 - y_i f(x_i))^2$
abs err loss: $|1 - y_i f(x_i)|$
logistic loss: $\log(1 + \exp(-y_i f(x_i)))$
hinge: $\max\{0, 1 - y_i f(x_i)\}$

14.4 Linear Models

$$\begin{aligned} f(x_i) &= w^T x_i \\ &= \sum_{j=1}^n w_j x_{ij} \end{aligned}$$

14.5 GLMs

Parametric model $p(y_i|x_i) = p(y_i|w^T x_i)$
Binary Classification $p(y|x)$ is Bernoulli

$$\begin{aligned} y_i &\in \{-1, +1\} \\ p(y_i|w^T x_i) &= \frac{1}{1 + \exp(-y_i x_i^T w_i)} \\ \log(p(y_i|w^T x_i)) &= -\log(1 + \exp(-y_i x_i^T w_i)) \end{aligned}$$

MLE of w

$$\begin{aligned} &\max_w \prod_{i=1}^n p(y_i|w^T x_i) \\ &\Rightarrow \min_w - \sum_{i=1}^n \log p(y_i|x_i^T w) \\ &\Rightarrow \min_w \sum_{i=1}^n \log(1 + \exp(-y_i x_i^T w_i)) \end{aligned}$$

Example 9. $X|Y \sim N\left(y\frac{w}{2}, I\right)$

$$y \in \{-1, +1\}, p(y = +1) = p(y = -1)$$

opt classifier

$$
\begin{aligned}
p\left(y|x\right) &= \frac{p\left(x|y\right)p\left(y\right)}{p\left(x\right)}\\
p\left(x\right) &= p\left(x,y=+1\right)+p\left(x,y=-1\right)\\
p\left(x|y=+1\right) &= \exp\left(-\frac{1}{2}\left(x-\frac{w}{2}\right)^T\left(x-\frac{w}{2}\right)\right)\\
&= \exp\left(-\frac{1}{2}x^Tx\right)\exp\left(\frac{1}{8}w^Tw\right)\exp\left(\frac{1}{2}w^Tx\right)
\end{aligned}
$$

$$
\begin{aligned}
p\left(y=+1|X=x\right) &= \frac{p\left(x|y=+1\right)}{p\left(x|y=+1\right)+p\left(x|y=-1\right)}\\
&= \frac{\exp\left(\frac{1}{2}w^Tx\right)}{\exp\left(\frac{1}{2}w^Tx\right)+\exp\left(-\frac{1}{2}w^Tx\right)}\\
&= \frac{1}{1+\exp\left(-w^Tx\right)}
\end{aligned}
$$

MLE of μ

$$
\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}y_ix_i\\
\mathbb{E}\left[\hat{\mu}\right] &= \mu = \frac{w}{2}\\
\Rightarrow \hat{w} &= 2\hat{\mu}
\end{aligned}
$$

15 Lecture 15

15.1 Optimization in ML

$$
\begin{aligned}
&\min_{w\in\mathbb{R}^d}\sum_{t=1}^{T}f_t\left(w\right)\\
f_t\left(w\right) &= \left(y_t-w^Tx_t\right)^2\\
f_t\left(w\right) &= \log\left(1-\exp\left(-y_tw^Tx_t\right)\right)\\
f_t\left(w\right) &= \max\left(0,1-y_tx_t^Tw\right)
\end{aligned}
$$

$$
\begin{aligned}
\left(x_t,y_t\right) &\overset{unif}{\sim}\left\{\left(x_i,y_i\right)\right\}_{i=1}^{n}\\
\frac{1}{T}\sum_{t=1}^{T}f_t\left(w\right) &\to^{as}\frac{1}{n}\sum_{i=1}^{n}f_i\left(w\right)
\end{aligned}
$$

15.2 Prob Model Approach

$$
\begin{aligned}
& p\left(y|w^T x\right) \\
& \min_w \sum_{t=1}^{n} \underbrace{-\log p\left(y_t|w^T x_t\right)}_{f_t(w)} \\
& \Rightarrow p\left(y_t|w^T x_t\right) = \exp\left(-f_t\left(w\right)\right)
\end{aligned}
$$

What properties should f_t have?

1. f_t is convex function (opt)
2. f_t is non-negative (prob interpretation)

$$
\begin{aligned}
& f_t : \mathbb{R} \to [0, \infty) \\
& \exp\left(-f_t\right) \in [0, 1]
\end{aligned}
$$

SGD $\{(x_i, y_i)\}_{i=1}^n$

$$
\begin{aligned}
(x_t, y_t) \overset{unif}{\sim} \{(x_i, y_i)\}_{i=1}^n, t = 1, 2, \ldots \\
w_1 \in \mathbb{R}^d \text{ arbitrary} \\
w_{t+1} = w_t - \gamma_t \nabla f_t\left(w_t\right), t = 1, 2, \ldots
\end{aligned}
$$

γ_t is stop size

Key fact

If f_t is convex:

$$
f_t\left(w^\star\right) \geqslant f_t\left(w\right) + \left(w^\star - w\right)^T \nabla f_t\left(w\right)
$$

Theorem 9. *$\gamma_t = \gamma$ fixed, $\|\nabla f_t\left(w\right)\| \leqslant G \; \forall \; t, w$*

Let $w_T^\star = \arg\min_w \sum_{t=1}^{T} f_t\left(w\right)$

$$
\frac{1}{T} \sum_{t=1}^{T} f_t\left(w_t\right) - f_t\left(w_T^\star\right) \leqslant \frac{\|w_1 - w^\star\|^2}{2\gamma T} + \frac{\gamma}{2} G^2
$$

Remark: choose γ so,

$$
\begin{aligned}
\frac{1}{\gamma T} \approx \gamma \Rightarrow \gamma = \frac{1}{\sqrt{T}} \\
O\left(\frac{1}{\sqrt{T}}\right)
\end{aligned}
$$

Proof. :

$$
\|w_{t+1} - w^\star\|^2 = \|w_t - \gamma \nabla f_t\left(w_t\right) - w^\star\|^2
$$

$$
\begin{aligned}
&= \|w_t - w^\star\|^2 - 2\gamma\left(w_t - w^\star\right)^T \nabla f_t\left(w_t\right) + \gamma^2 \underbrace{\nabla f_t^T\left(f_t\right)\nabla f_t\left(w_t\right)}_{\|\nabla f_t(w_t)\|^2 \leqslant G^2} \\
&\Rightarrow \left(w_t - w^\star\right)^T \nabla f_t\left(w_t\right) \leqslant \frac{\|w_t = w^\star\|^2 - \|w_{t+1} - w^\star\|^2}{2\gamma} + \frac{\gamma G^2}{2} \\
&\Rightarrow f_t\left(w_t\right) - f_t\left(w^\star\right) \leqslant \left(w_t - w^\star\right)^T \nabla f_t\left(w_t\right) \leqslant \frac{\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2}{2\gamma} + \frac{\gamma G^2}{2} \\
&\Rightarrow \frac{1}{T}\sum_{t=1}^{T} f_t\left(w_t\right) - f_t\left(w^\star\right) \leqslant \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2}{2\gamma} + \frac{\gamma G^2}{2}\right) \\
&\Rightarrow \frac{1}{T}\sum_{t=1}^{T} f_t\left(w_t\right) - f_t\left(w^\star\right) \leqslant \frac{\|w_1 - w^\star\|^2}{2\gamma} + \frac{\gamma G^2}{2}
\end{aligned}
$$

where,

$$
\begin{aligned}
&\sum_{t=1}^{T} \|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2 \\
&= \|w_1 - w^\star\|^2 - \|w_2 - w^\star\|^2 + \|w_2 - w^\star\|^2 - \|w_3 - w^\star\|^2 + \ldots - \|w_{T+1} - w^\star\|^2 \\
&= \|w_1 - w^\star\|^2 - \|w_{T+1} - w^\star\|^2 \\
&\leqslant \|w_1 - w^\star\|^2
\end{aligned}
$$

□

Theorem 10. $\gamma_t = \dfrac{1}{\sqrt{t}}, \|w_t\| \leqslant B, \|\nabla f_t\left(w\right)\| \leqslant G \ \forall\ t, w$

Let $w_T^\star = \arg\min\limits_{w} \sum\limits_{t=1}^{T} f_t\left(w\right)$

$$
\frac{1}{T}\sum_{t=1}^{T} f_t\left(w_t\right) - f_t\left(w_T^\star\right) \leqslant \frac{2B^2 + G^2}{\sqrt{T}} \ \forall\ T
$$

Proof. Similar,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} f_t\left(w_t\right) - f_t\left(w^\star\right) &\leqslant \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2}{2\gamma_t} + \frac{\gamma_t G^2}{2}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{t}\frac{\|w_t - w^\star\|^2}{2} - \sqrt{t}\frac{\|w_{t+1} - w^\star\|^2}{2} + \frac{1}{2\sqrt{t}}G^2\right) \\
&= \frac{1}{T}\left(\frac{\|w_1 - w^\star\|^2}{2} - \frac{\|w_{T+1} - w^\star\|^2}{2}\right) + \frac{1}{T}\sum_{t=2}^{T}\frac{\|w_t - w^\star\|^2}{2}\left(\sqrt{t} - \sqrt{t-1}\right) + \frac{1}{T}\sum_{t=1}^{T}\frac{G^2}{2\sqrt{t}} \\
&\leqslant \frac{1}{T}\left(\frac{\|w_1 - w^\star\|^2}{2} - \frac{\|w_{T+1} - w^\star\|^2}{2}\right) + \frac{1}{T}\sum_{t=2}^{T} 2B^2\left(\sqrt{t} - \sqrt{t-1}\right) + \frac{1}{T}\sum_{t=1}^{T}\frac{G^2}{2\sqrt{t}} \\
&\leqslant \frac{2B^2}{T} + \frac{2B^2}{\sqrt{T}} + \frac{G^2}{\sqrt{T}}
\end{aligned}
$$

where,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \approx \int_{1}^{T} \frac{1}{\sqrt{t}} dt \leqslant \sqrt{T}$$

$\square$

16 Lecture 16

16.1 Regularization and Bayesian Inference

MLE GLE: $p\left(y_i | w^T x_i\right)$

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} -\log p\left(y_i | w^T x_i\right)$$

Gaussian

$$\begin{aligned}
& y_i | w^T x_i \sim N\left(w^T x_i, I\right) \\
& \min_{w} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \\
& \min_{w} \|y - Xw\|^2 \\
& X^T y = X^T X \hat{w}
\end{aligned}$$

system of n equations in d unknowns

$$\{(x_i, y_i)\}_{i=1}^{n}, n < d$$

underdetermined
if $n < d$,
then exists $v \in \mathbb{R}^d$
such that

$$\begin{aligned}
Xv &= 0 \text{ and } v \neq 0 \\
x_i^T v &= 0, i = 1, ..., n \\
X &= \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, v = \begin{bmatrix} \alpha \\ -\alpha \\ \alpha \end{bmatrix} \\
d &= 3, n = 2 \\
\left(\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, 1\right), \left(\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, -1\right), Xv &= 0 \; \forall \; \alpha \in \mathbb{R} \\
X^T y &= X^T X \left(\hat{w} + v\right) = X^T X \hat{w}
\end{aligned}$$

$\hat{w}$ is a solution
then so is $\hat{w}+v$
What solution should we use?

$$\hat{w}=\arg\min_{w:x^Ty=x^Txw}\|w\|$$

minimum norm solution

$$\begin{aligned}&(x,y)\,,x=x_i+\varepsilon\\&w^Tx=w^Tx_i+w^T\varepsilon\leqslant\|w\|\|\varepsilon\|\end{aligned}$$

16.2 Computing Minimum Norm Solution

1. $\min\limits_{w}\|w\|^2$ such that $X^Ty=X^TXw$
2. $\min\limits_{w}\|y-Xw\|^2+\tau\|w\|^2,$ tiny $\tau>0$
3. $X^TX=U\Lambda U^T,\Lambda=$ diag $(\lambda_1,\lambda_2,...,0,0)\,,\left(X^TX\right)^{-1}=U\Lambda^{-1}U^T,\hat{w}=\left(X^TX\right)^{-1}X^Ty$

$$\begin{aligned}\|y-Xw\|^2&=\sum\left(y_i-x_i^Tw\right)^2\\&=-\sum\log p\left(y_i|w^Tx_i\right)\\\tau\|w\|^2&=\tau\sum_{j=1}^{d}w_j^2\\&=-\sum_{j=1}^{d}\log p\left(w_j\right)\\&\Rightarrow-\log p\left(w_j\right)\propto\tau w_j^2\\p\left(w_j\right)&=\exp\left(-\log p\left(w_j\right)\right)\propto\exp\left(-\tau w_j^2\right)\end{aligned}$$

prior
Likelihood

$$\begin{aligned}y|Xw\sim N\left(Xw,I\right)&=p\left(y|w\right)\\w\sim N\left(0,\frac{1}{2\tau}I\right)&=p\left(w\right)\end{aligned}$$

prior prob for w

$$\begin{aligned}&\min_{w}-\log p\left(y|w\right)-\log p\left(w\right)\\&=\max_{w}p\left(y|w\right)p\left(w\right)\end{aligned}$$

$$
\begin{aligned}
&= \max_w \frac{p(y|w)\,p(w)}{p(y)} \\
&= \max_w p(w|y)
\end{aligned}
$$

posterior distribution of w given y
Maximum a Posteriori estimator (MAP)

16.3 Bayesian Inference

$$
\begin{aligned}
&x, \{p(x|\theta)\}_{\theta\in\Theta} \\
&x_1, x_2, ..., x_n \overset{iid}{\sim} q
\end{aligned}
$$

MLE:

$$
\max_{\theta\in\Theta} \prod_{i=1}^{n} p(x_i|\theta)
$$

$\hat{\theta}$ is MLE
roughly, $p_{\hat{\theta}}$ is density that approximately

$$
\min_\theta \text{ KL } \left(q, p_{\hat{\theta}}\right)
$$

We know or believe something a priori about θ
$p(\theta)$ is a weighting function on the Θ

$$
\theta_{MAP} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} p(x_i|\theta)\,p(\theta)
$$

Example 10. $x_i \sim \text{Poisson}(\theta), \theta > 0$

$$
p(x_i = k|\theta) = e^{-\theta}\frac{\theta^k}{k!}, k = 1, 2, ...
$$

x_i = number of times a "word" appears index i

$$
\mathbb{E}[x_i] = \theta
$$

MLE:

$$
\begin{aligned}
&\min_\theta \underbrace{\sum_{i=1}^{n} (\theta - x_i \log\theta)}_{L(\theta)} \\
&\frac{\partial}{\partial\theta} L(\theta) = n - \frac{\sum x_i}{\theta} \\
&\hat{\theta}_{MLE} = \frac{1}{n}\sum x_i
\end{aligned}
$$

unbiased
MAP:

$$p\left(\theta\right)=\alpha e^{-\alpha\theta}$$

exponential distribution

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta}\prod_{i}p\left(x_i|\theta\right)p\left(\theta\right)\\ &= \arg\min_{\theta}\sum_{i}\left(\theta - x_i\log\theta+\alpha\theta\right)\\ &\Rightarrow \sum_{i}\left(\left(1+\alpha\right)\theta - x_i\log\theta\right)\\ &\Rightarrow \hat{\theta}_{MAP}=\frac{1}{\left(1+\alpha\right)n}\sum x_i=\frac{1}{1+\alpha}\hat{\theta}_{MLE}\end{aligned}$$

iid Poiss(θ)

$$\begin{aligned}\sum_{x_i},\mathbb{E}\left[\sum x_i\right] &= n\theta\\ \mathbb{V}\left[\sum x_i\right] &= \sum\mathbb{V}\left[X_i\right]=n\theta\\ \mathbb{V}\left[\frac{1}{n}\sum x_i\right] &= \frac{\theta}{n}\end{aligned}$$

For MLE,

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_{MLE}\right] &= \theta\\ \mathbb{V}\left[\hat{\theta}_{MLE}\right) &= \frac{\theta}{n}\end{aligned}$$

For MAP,

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_{MAP}\right] &= \frac{1}{1+\alpha}\theta\\ \mathbb{V}\left[\hat{\theta}_{MAP}\right] &= \left(\frac{1}{1+\alpha}\right)^2\frac{\theta}{n}\end{aligned}$$

MSE

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \text{ Bias }^2+\text{ Variance }\\ &= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]+\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2\right]\\ &= \left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2+\mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)^2\right]\end{aligned}$$

$$\text{MSE }\left[\hat{\theta}_{MLE}\right]=0+\frac{\theta}{n}=\frac{\theta}{n}$$

$$\text{MSE}\left[\hat{\theta}_{MAP}\right]=\left(1-\frac{1}{1+\alpha}\right)^2\theta^2+\left(\frac{1}{1+\alpha}\right)^2\frac{\theta}{n}$$

$$\alpha=1\Rightarrow\frac{\theta^2}{4}+\frac{1}{4}\frac{\theta}{n}=\frac{\theta}{4}\left(\theta+\frac{1}{n}\right)$$

$$\frac{1}{4}\left(\theta+\frac{1}{n}\right)\gtreqless\frac{1}{n}$$

17 Lecture 17

17.1 Bayesian Inference

prior distribution: $p\left(\theta\right)$
likelihood: $p\left(x|\theta\right)$
posterior distribution: $p\left(\theta|x\right)\propto p\left(x|\theta\right)p\left(\theta\right)$

$$\begin{aligned}\hat{\theta}_{MAP}&=\arg\max_{\theta\in\Theta}p\left(x|\theta\right)p\left(\theta\right)\\&=\arg\min_{\theta}\left(-\log p\left(x|\theta\right)-\log p\left(\theta\right)\right)\end{aligned}$$

17.2 Linear Bayesian Regression

$$\begin{bmatrix}y_1\\ \ldots\\ y_n\end{bmatrix}=y=Xw+e$$

$$e\sim N\left(0,\sigma^2I\right)$$

$$\hat{\theta}_{LS,MLE}=\left(X^TX\right)^{-1}X^Ty$$

$$y|w\sim N\left(Xw,\sigma^2I\right)$$

$$p\left(y|w\right)\propto\exp\left(-\frac{1}{2\sigma^2}\|y-Xw\|^2\right)$$

prior:

$$w\sim N\left(0,\sigma_w^2I\right)$$

$$\mathbb{E}\left[w\right]=0$$

$$\mathbb{E}\left[\|w\|^2\right]=\sigma_w^2d$$

$$p\left(w\right)\propto\exp\left(-\frac{1}{2\sigma_w^2}\|w\|^2\right)$$

$$\begin{aligned}\hat{w}_{MAP}&=\arg\max_{w}p\left(y|w\right)p\left(w\right)\\&=\arg\max_{w}\exp\left(-\frac{1}{2\sigma^2}\|y-Xw\|^2\right)\exp\left(-\frac{1}{2\sigma_w^2}\|w\|^2\right)\\&=\arg\min_{w}\frac{1}{2\sigma^2}\|y-Xw\|^2+\frac{1}{2\sigma_w^2}\|w\|^2\end{aligned}$$

$$
\begin{aligned}
&= \arg\min_w \frac{1}{\sigma^2}\left(\underbrace{\frac{1}{2}\|y-Xw\|^2+\frac{1}{2}\lambda\|w\|^2}_{f_\lambda(w)}\right), \lambda=\frac{\sigma^2}{\sigma_w^2}>0 \\
0 &= \frac{\partial f_\lambda(w)}{\partial w} = -X^T(y-Xw)+\lambda w \\
X^T y &= X^T X w + \lambda w \\
&= \left(X^T X + \lambda I\right) w \\
\hat{w}_{MAP} &= \left(X^T X + \lambda I\right)^{-1} X^T y
\end{aligned}
$$

ridge regression estimate

$$
\begin{aligned}
y &= w + e \\
y|w &\sim N\left(w,\sigma^2\right) \\
w &\sim N\left(0,\sigma_w^2\right) \\
w|y &\sim N\left(\frac{\sigma_w^2}{\sigma^2+\sigma_w^2}y, \frac{\sigma^2\sigma_w^2}{\sigma^2+\sigma_w^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
p(w|y) &\propto p(y|w)\,p(w) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(y-w)^2-\frac{1}{2\sigma_w^2}w^2\right) \\
&\propto \exp\left(-\frac{\sigma_w^2(y-w)^2+\sigma^2 w^2}{2\sigma^2\sigma_w^2}\right) \\
&= \exp\left(-\frac{w^2-2\frac{\sigma_w^2}{\sigma^2+\sigma_w^2}w+\frac{\sigma_w^2}{\sigma^2+\sigma_w^2}y^2}{\frac{2\sigma^2\sigma_w^2}{\sigma^2+\sigma_w^2}}\right) \\
&\propto \exp\left(-\frac{\left(w-\frac{\sigma_w^2}{\sigma^2+\sigma_w^2}y\right)^2}{2\frac{\sigma^2\sigma_w^2}{\sigma^2+\sigma_w^2}}\right)
\end{aligned}
$$

$$
\begin{aligned}
\hat{w}_{MAP} &= \frac{\sigma_w^2}{\sigma^2+\sigma_w^2}y = \frac{\sigma_w^2}{\sigma^2+\sigma_w^2}\underbrace{\hat{w}_{MLE}}_{y} \\
\frac{\sigma_w^2}{\sigma^2+\sigma_w^2} &= \frac{1}{1+\frac{1}{\lambda}} \\
\lambda &= \frac{\sigma_w^2}{\sigma^2} = \text{ SNR}
\end{aligned}
$$

signal to noise ratio

17.3 Gauss Markov Theorem

If x, y are jointly Gaussian vector

$$
\begin{aligned}
\begin{bmatrix} x \\ y \end{bmatrix} &\sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right) \\
y|x &\sim N\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(x - \mu_x\right), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right) \\
\mathbb{E}\left[\left(x - \mu_x\right)\left(y - \mu_y\right)^T\right] &= \Sigma_{xy}
\end{aligned}
$$

Example 11. $y = Xw + e, e \sim N\left(0, \sigma^2 I\right)$

$$
\begin{aligned}
\Sigma_{ww} &= \mathbb{E}\left[ww^T\right] = \sigma_w^2 \\
\Sigma_{wy} &= \mathbb{E}\left[w\left(Xw + e\right)^T\right] \\
&= \mathbb{E}\left[ww^T\right]X^T \\
&= \sigma_w^2 X^T \\
\Sigma_{yy} &= \mathbb{E}\left[\left(Xw + e\right)\left(Xw + e\right)^T\right] \\
&= XX^T\sigma_w^2 + \sigma^2 I \\
w &\sim N\left(0, \sigma_w^2 I\right) \\
\begin{bmatrix} y \\ w \end{bmatrix} &\sim N \\
\mathbb{E}\left[w|y\right] &= \sigma_w^2 X^T\left(XX^T + \sigma^2 I\right)^{-1} y \\
\hat{w}_{MAP} &= \left(X^T X + \frac{\sigma_w^2}{\sigma^2} I\right)^{-1} X^T y
\end{aligned}
$$

17.4 Weiner Filter

$$
X = By + e, e \sim N\left(0, \sigma^2 I\right)
$$

y is an image
B is blur
e is noise
X blurry noise image
Likelihood,

$$
\begin{aligned}
X|y &\sim N\left(By, \sigma^2 I\right) \\
\hat{y}_{MLE} &= \arg\max_y \frac{1}{2}\|x - By\|^2 = B^{-1}X
\end{aligned}
$$

if B^{-1} exist,

$$
\hat{y}_{MLE} = B^{-1}x = y + B^{-1}e
$$

blow up noise

$$
\begin{aligned}
y &\sim N\left(0, \Sigma_{yy}\right) \\
p\left(y|x\right) &\propto \exp\left(-\frac{1}{2\sigma^2}\|x-By\|^2\right)\exp\left(-\frac{1}{2}y^T\Sigma_{yy}^{-1}y\right) \\
\hat{y}_{MAP} &= \mathbb{E}\left[y|x\right] = \arg\min_y \underbrace{\frac{1}{2\sigma^2}\|x-By\|^2+\frac{1}{2}y^T\Sigma_{yy}^{-1}y}_{f(y)} \\
0 &= \frac{\partial f\left(y\right)}{\partial y} = \frac{1}{\sigma^2}\left(-B^T\left(X-By\right)+\sigma^2\Sigma_{yy}^{-1}y\right) \\
0 &= -B^TX+B^TBy+\sigma^2\Sigma_{yy}^{-1}y \\
B^TX &= \left(B^TB+\sigma^2\Sigma_{yy}^{-1}\right)y \\
\hat{y}_{MAP} &= \left(B^TB+\sigma^2\Sigma_{yy}^{-1}\right)^{-1}B^TX
\end{aligned}
$$

If $\sigma^2 << 1$,

$$
\hat{y}_{MAP} = \left(B^TB\right)^{-1}B^TX = B^{-1}B^{-T}B^TX = B^{-1}X
$$

17.5 Linear Regression

$$
\begin{aligned}
y &= Xw+e, e\sim N\left(0,\sigma^2I\right) \\
y|w &\sim N\left(Xw,\sigma^2I\right)
\end{aligned}
$$

Ridge Prior

$$
\begin{aligned}
p\left(w\right) &\propto \exp\left(-\lambda\|w\|^2\right) \\
&\Rightarrow \hat{w}_{MAP} = \left(X^TX+\lambda I\right)^{-1}X^Ty
\end{aligned}
$$

Prior knowledge: most w_j are $0, j=1,...,d$
i.e. most features not important

$$
\begin{aligned}
&\sum_{j=1}^{d}\mathbb{1}_{\{w_j\neq 0\}} \text{ is small } << d \\
p\left(w\right) &\propto \exp\left(-\lambda\sum_{j=1}^{d}\mathbb{1}_{\{w_j\neq 0\}}\right), \lambda>0 \\
p\left(w|y\right) &\propto \exp\left(-\frac{1}{2\sigma^2}\|y-Xw\|^2\right)\exp\left(-\lambda\sum\mathbb{1}_{\{w_j\neq 0\}}\right) \\
\hat{w}_{MAP} &= \arg\min_w\left(\frac{1}{2\sigma^2}\|y-Xw\|^2+\lambda\sum_{i=1}^{d}\mathbb{1}_{\{y_j\neq 0\}}\right)
\end{aligned}
$$

hard to optimize

$$p(w) \propto \exp\left(-\lambda \sum_{j=1}^{d} |w_j|\right)$$

$$\hat{w}_{MAP} = \arg\min_{w} \left(\frac{1}{2\sigma^2}\|y - Xw\|^2 + \lambda\|w\|_1\right)$$

easy to optimize
LASSO

18 Lecture 18

18.1 Linear Prediction

$$\hat{y} = w^T x$$

$$\min_{w} \sum_{i=1}^{n} \underbrace{l\left(y_i, w^T x_i\right)}_{-\log like} + \underbrace{\lambda \text{ pen } (w)}_{-\log prior}$$

$$p(y|w) \propto \exp\left(-l\left(y_i, w^T x_i\right)\right)$$

18.2 Loss Functions

sq err: $\|y - Xw\|_2^2 = \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2$

abs err: $\|y - Xw\|_1 = \sum_{i=1}^{n} |y_i - w^T x_i|$

$0 - 1$ loss: $\sum_{i=1}^{n} \mathbb{1}_{\{y_i - w^T x_i < 0\}}$

logistic loss: $\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right)$

hinge: $\sum_{i=1}^{n} \max\left\{0, 1 - y_i w^T x_i\right\}$

18.3 Penalties, Regularizers

2 norm: $\|w\|^2$

1 norm: $\|w\|_1 = \sum_{j=1}^{d} |w_j|$

ideal pen: $\sum_{j=1}^{d} \mathbb{1}_{\{w_j \neq 0\}}$

LASSO

$$\min_{w} \frac{1}{2}\|y - Xw\|^2 + \lambda\|w\|_1$$

$$d = n, X = I$$

$$
\begin{aligned}
&\min_w \sum_{i=1}^n \frac{1}{2}(y_i - w_i)^2 + \lambda |w_i| \\
&\min_{w_i} \underbrace{\frac{1}{2}(y_i - w_i)^2 + \lambda |w_i|}_{f(w_i)} \\
&0 = \frac{\partial f}{\partial w_i} = -y_i + w_i + \lambda \operatorname{sign}(w_i), w_i \neq 0 \\
&w_i = y_i - \lambda \operatorname{sign}(w_i) \\
&\Rightarrow \operatorname{sign}(w_i) = \operatorname{sign}(y_i)
\end{aligned}
$$

$$
\hat{w}_i = \begin{cases} y_i - \lambda \operatorname{sign}(y_i) \\ 0 \end{cases}
$$

$$
f(\hat{w}_i) = \begin{cases} \frac{\lambda^2}{2} + \lambda |y_i - \lambda \operatorname{sign}(y_i)| \\ \frac{y_i^2}{2} \end{cases}
$$

$$
\begin{aligned}
&|y_i| < \lambda \Rightarrow \hat{w}_i = 0 \\
&|y_i| > \lambda \Rightarrow \hat{w}_i = y_i - \lambda \operatorname{sign}(y_i) = \operatorname{sign}(y_i)(|y_i| - \lambda)
\end{aligned}
$$

soft threshold function

w_i unknown, $\varepsilon_i \sim N(0,1)$

$$
\begin{aligned}
&y_i \sim N(w_i, 1) \\
&y_i = w_i + \varepsilon_i
\end{aligned}
$$

1. MLE

$$
\begin{aligned}
&\min_w \frac{1}{2}\|y - w\|^2 \\
&\Rightarrow \hat{w}_i = y_i \\
&\mathbb{E}\left[\|\hat{w} - w\|^2\right] = n
\end{aligned}
$$

2. G-MAP

$$
\begin{aligned}
&\min_w \frac{1}{2}\|y - w\|^2 + \lambda \|w\|^2 \\
&\hat{w}_i = \frac{1}{1+\lambda} y_i \\
&\mathbb{E}\left[\|\hat{w} - w\|^2\right] = \left(\frac{\lambda}{1+\lambda}\right)^2 \|w\|^2 + \left(\frac{1}{1+\lambda}\right)^2 n
\end{aligned}
$$

3. Soft threshold

$$
\hat{w}_i = \operatorname{sign}(y_i)(|y_i| - \lambda)_+
$$

4. Oracle, $\sigma^2 = 1$

$$\hat{w}_i = \begin{cases} 0 & \text{if } |w_i| \leqslant 1 \\ y_i & \text{if } |w_i| > 1 \end{cases}$$

$$\mathbb{E}\left[\|\hat{w}_o - w\|^2\right] = \sum_{i=1}^{n} \min\left\{w_i^2, 1\right\}$$

Theorem 11. *Assume $y_i \sim N(w_i, 1)$, $\hat{w}_i = sign(y_i)(|y_i| - \lambda)_+$, and take $\lambda = \sqrt{2\log n}$, then*

$$\mathbb{E}\left[\|\hat{w} - w\|^2\right] \leqslant (2\log n + 1)\left(1 + \sum_{i=1}^{n} \min\left\{w_i^2, 1\right\}\right)$$

Example 12. $k < n, w_i \neq 0$, then,

$$\sum_{i=1}^{n} \min\left\{w_i^2, 1\right\} = k$$
$$(2\log n + 1)\left(1 + \sum_{i=1}^{n} \min\left\{w_i^2, 1\right\}\right) = O(k\log n)$$

$$\begin{aligned} y &\sim N(0,1) \\ \mathbb{P}\{y \geqslant \lambda\} &= \frac{1}{\sqrt{2\pi}}\int_{\lambda}^{\infty} e^{-\dfrac{y^2}{2}}\,dy \\ &\leqslant \frac{1}{2}e^{-\dfrac{\lambda^2}{2}} \\ \mathbb{P}\{|y_i| > \lambda\} &\leqslant e^{-\dfrac{\lambda^2}{2}} \\ \mathbb{P}\{|y_{i,1}| \geqslant \lambda \text{ or } |y_{i,2}| \geqslant \lambda \text{ or } ...|y_{i,n}| \geqslant \lambda\} &\leqslant ne^{-\dfrac{\lambda^2}{2}} \\ &= e^{-\dfrac{1}{2}\left(\lambda^2 - 2\log n\right)} \end{aligned}$$

union bound

19 Lecture 19

$$\begin{aligned} &\min_w \|y - Xw\|^2 + \lambda \underbrace{pen(w)}_{\|w\|^2 \text{ or } \|w\|_1} \\ &\min_w (y - w)^2 + \lambda pen(w) \end{aligned}$$

has a closed-form solution

$$\hat{w}_i = y_i - \text{sign}(y_i)\min\{|y_i|, \lambda\}$$
$$L(w) = \|y - Xw\|^2 + \lambda pen(w)$$

iterates $w_1, w_2, ...$

$$\begin{aligned}
L(w_1) &\geqslant L(w_2) \geqslant L(w_3)\ldots \\
L(w) &= \|y - Xw_k + Xw_k - Xw\|^2 + \lambda pen(w) \\
&= \|y - Xw_k\|^2 + \underbrace{\|X(w_k - w)\|^2 + 2(y - Xw_k)^T X(w_k - w) + \lambda pen(w)}_{\text{choose } w \text{ so this } \boxed{x} \leqslant \lambda pen(w_k)} \\
w_{k+1} &= \arg\min_w \boxed{x} \\
\boxed{x} &= \|X(w_k - w)\|^2 + 2(y - Xw_k)^T X(w_k - w) + \lambda pen(w) \\
&\leqslant \underbrace{\|X\|^2}_{\leqslant \gamma^{-1}} \|w_k - w\|^2 + 2\underbrace{(y - Xw_k)^T X}_{\frac{1}{\gamma}V_k^T}(w_k - w) + \lambda pen(w), V_k = \gamma X^T(y - Xw_k) \\
\gamma\boxed{x} &\leqslant \|w_k - w\|^2 + 2V_k^T(w_k - w) + \gamma\lambda pen(w) \\
&= \|V_k + w_k - w\|^2 - \|V_k\|^2 + \gamma\lambda pen(w)
\end{aligned}$$

then the minimization problem is

$$\min_w \|V_k + w_k - w\|^2 + \gamma\lambda pen(w)$$
$$z_k = w_k + \gamma X^T(y - Xw_k)$$

GD iterate

$$\min_w \|z_k - w\|^2 + \gamma\lambda pen(w)$$
$$\min_w \sum_{j=1}^{d}\left(\left(z_{k_j} - w_j\right)^2 + \gamma\lambda|w_j|\right).$$

Goal:

$$\min_w \|y - Xw\|^2 + \lambda\|w\|_1$$

w_1 init

$$\begin{aligned}
k &= 1, 2, ... \\
z_k &= w_k + \gamma X^T(y - Xw_k) \\
\hat{w}_{k+1} &= z_k - \text{sign}(z_k)\min\{|z_k|, \gamma\lambda\} \\
\min_w L(w) &= \|y - Xw\|^2 + \lambda\|w\|^2 \\
\hat{w} &= \left(X^T X + \lambda I\right)^{-1} X^T y
\end{aligned}$$

$$
\begin{aligned}
X_{n\times d}, w &\in \mathbb{R}^d \\
\frac{\partial L}{\partial w} &= -2X^T\left(y - Xw\right) + 2\lambda w = 0 \\
X^T\left(y - Xw\right) &= \lambda w
\end{aligned}
$$

solution have form

$$
w = \frac{1}{\lambda}X^T\underbrace{\left(y - Xw\right)}_{\lambda\alpha}
$$

form of solution

$$
\begin{aligned}
w &= X^T\alpha, \alpha \in \mathbb{R}^n \\
X &= \begin{bmatrix} x_1^T \\ x_2^T \\ \dots \\ x_n^T \end{bmatrix} \\
X^T &= \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \\
w &= \sum_{i=1}^{n}\alpha_i x_i \text{ for } \alpha_i \in \mathbb{R}
\end{aligned}
$$

weights are linear combo of features

19.1 Dual Opt

$$
\begin{aligned}
\min_{\alpha\in\mathbb{R}^n} &\|y - XX^T\alpha\|^2 + \lambda\|X^T\alpha\|^2 \\
L\left(\alpha\right) &= \|y - XX^T\alpha\|^2 + \lambda\alpha^TXX^T\alpha \\
\frac{\partial L}{\partial \alpha} &= 2XX^T\left(y - XX^T\alpha\right) + 2\lambda XX^T\alpha \\
&= 2XX^T\left(-y + XX^T\alpha + \lambda\alpha\right) = 0 \\
y &= \left(XX^T + \lambda I\right)\alpha \\
\hat{\alpha} &= \left(K + \lambda I\right)^{-1}y \\
\hat{w} &= X^T\hat{\alpha} \\
K &= XX^T
\end{aligned}
$$

kernel matrix

$$
\begin{aligned}
K_{ij} &= K\left(X_i, X_j\right) = X_i^TX_j \\
\hat{w} &= \left(X^TX + \lambda I\right)^{-1}X^Ty \\
&= X^T\left(XX^T + \lambda I\right)^{-1}y \\
X &= UDV^T
\end{aligned}
$$

19.2 Using nonliear features

$$
\begin{aligned}
X &= \begin{bmatrix} x_1 \\ ... \\ x_d \end{bmatrix} \in \mathbb{R}^d \\
\Phi\left(X\right) &= \begin{bmatrix} \phi_1\left(x\right) \\ ... \\ \phi_D\left(x\right) \end{bmatrix} \in \mathbb{R}^D \\
\Phi &= \begin{bmatrix} \Phi\left(x_1\right)^T \\ ... \\ \Phi\left(x_n\right)^T \end{bmatrix} \\
&\min_w \|y - \Phi w\|^2 + \lambda \|w\|^2 \\
\hat{w} &= \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T y \\
\hat{\alpha} &= \left(\underbrace{\Phi \Phi^T}_{K_\Phi} + \lambda \Phi\right)^{-1} y \\
K_\Phi\left(i, j\right) &= \Phi\left(x_i\right)^T \Phi\left(x_j\right)
\end{aligned}
$$

Example 13. $x \in \mathbb{R}^2$

$$
\begin{aligned}
\Phi\left(x\right) &= \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix} \in \mathbb{R}^3 \\
K_\Phi\left(i, j\right) &= \Phi\left(x_i\right)^T \Phi\left(x_j\right) \\
&= x_{i1}^2 x_{j1}^2 + 2 x_{i1} x_{i2} x_{j1} x_{j2} + x_{i2}^2 x_{j2}^2 \\
&= \left(x_{i1} x_{j1} + x_{i2} x_{j2}\right)^2 \\
&= \left(x_i^T x_j\right)^2
\end{aligned}
$$

Example 14. Other kernels,

$$
\begin{aligned}
K_\Phi\left(i, j\right) &= \left(x_i^T x_j + 1\right)^2 \\
K_\Phi\left(i, j\right) &= \exp\left(-\lambda \|x_i - x_j\|^2\right) \\
\hat{y} &= \sum_{i=1}^{n} \hat{\alpha}_i K_\Phi\left(x_i, x\right)
\end{aligned}
$$

20 Lecture 20

20.1 GM Thm

$$
\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)
$$

$$y|x \sim N\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(x-\mu_x\right), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$$

If $\mu_x = \mu_y = 0$, MAP is $\sigma_{xy}\sigma_{xx}^{-1}x$

20.2 Kernels

$$\min_{w\in\mathbb{R}^d} \|y - Xw\|^2 + \lambda\|w\|^2$$

$$X = \begin{bmatrix} x_1^T \\ ... \\ x_n^T \end{bmatrix}, n \times d$$

$$\phi : \mathbb{R}^d \to \mathbb{R}^D$$

$$\phi\left(x\right) = \begin{bmatrix} \phi_1\left(x\right) \\ ... \\ \phi_D\left(x\right) \end{bmatrix}$$

$$\min_{w\in\mathbb{R}^D} \|y - \Phi w\|^2 + \lambda\|w\|^2$$

$$\Phi = \begin{bmatrix} \phi\left(x_1\right)^T \\ ... \\ \phi\left(x_n\right)^T \end{bmatrix}, n \times D$$

$$\begin{aligned} \Phi^T\left(y - \Phi w\right) &= \lambda w \\ \hat{w} &= \Phi^T\alpha, \alpha \in \mathbb{R}^n \\ \hat{w} &= \sum_{i=1}^{n} \alpha_i \phi\left(x_i\right) \end{aligned}$$

$$\begin{aligned} \min_{\alpha} \|y - \Phi\left(\Phi^T\alpha\right)\|^2 + & \underbrace{\lambda\|\Phi^T\alpha\|^2}_{\lambda\alpha^T\Phi\Phi^T\alpha, K=\Phi\Phi^T} \\ \min_{\alpha} \|y - K\alpha\|^2 + & \lambda\alpha^T K\alpha \\ K_{ij} &= K\left(x_i, x_j\right) = \phi^T\left(x_i\right)\phi\left(x_j\right) \\ K &= \Phi\Phi^T \\ &= \begin{bmatrix} \phi^T\left(x_1\right) \\ ... \\ \phi^T\left(x_n\right) \end{bmatrix} \left[\phi^T\left(x_1\right)...\phi^T\left(x_n\right)\right] \\ K^T &= \left(\Phi\Phi^T\right)^T \\ &= \Phi\Phi^T \\ 2K\left(y - K\alpha\right) &= 2\lambda K\alpha \\ y - K\alpha &= \lambda\alpha \\ \hat{\alpha} &= \left(K + \lambda I\right)^{-1} y \end{aligned}$$

get new x, predict

$$\begin{aligned}
\hat{y} &= \hat{w}^T \phi(x) \\
&= \left(X^T \hat{\alpha}\right)^T \phi(x) \\
&= \left(\sum_{i=1}^{n} \hat{\alpha}_i \phi(x_i)\right)^T \phi(x) \\
&= \sum_{i=1}^{n} \hat{\alpha}_i \underbrace{\phi^T(x_i)\phi(x)}_{K(x_i,x)}
\end{aligned}$$

Example,

$$\begin{aligned}
x &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2, \phi(x) = \left[1, x_1, x_2, x_1^2, x_1x_2, x_2^2\right]^T \in \mathbb{R}^6 \\
\phi^T(x)\phi(z) &= 1 + x_1z_1 + x_2z_2 + x_1^2z_1^2 + x_1x_2z_1z_2 + x_2^2z_2^2 \\
\left(1 + x^Tz\right)^2 &= 1 + 2x^Tz + \left(x^Tz\right)^2 \\
&= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2 \\
K(x_i, x_j) &= \left(1 + x_i^Tx_j\right)^2 \\
\left(1 + x^Tz\right)^l &= \sum_{k=0}^{l} \binom{l}{k} (1)^{l-k} \left(x^Tz\right)^k \\
&= \sum_{k=0}^{l} \binom{l}{k} \left(x^Tz\right)^k \\
&D >> d \\
&D >> n
\end{aligned}$$

If $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\begin{aligned}
\phi(x) &= \{x_1^p x_2^q\}, p, q \geqslant 0, p + q \leqslant l \\
\phi(x) &= \{x_1^{p_1} x_2^{p_2} ... x_d^{p_d}\}, 0 \leqslant p_1 + p_2 + ... p_d \leqslant l
\end{aligned}$$

Gaussian kernel

$$\begin{aligned}
K(x_i, x_j) &= \exp\left(-\beta \|x_i - x_j\|^2\right) \\
&= \phi^T(x_i)\phi(x_j)
\end{aligned}$$

infinite dim kernel space

$$\begin{aligned}
e^x &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + ... \\
e^{-x^2} &= 1 - \frac{x^2}{1!} + \frac{x^4}{2!} \\
\hat{\alpha} &= (K + \lambda I)^{-1} y
\end{aligned}$$

”nonparametric” learning

21 Lecture 21

$$
\begin{aligned}
\{(x_i, y_i)\}_{i=1}^n, \phi : \mathbb{R}^d \to \mathbb{R}^D, x_i \in \mathbb{R}^d \\
\min_{w \in \mathbb{R}^D} \|y - \Phi w\|^2 + \lambda \|w\|^2 \\
\Phi = \begin{bmatrix} \phi^T(x_1) \\ \dots \\ \phi^T(x_n) \end{bmatrix} \\
\hat{y} = \phi(x_i)^T w \\
= w^T \phi(x_i) \\
\Phi^T (y - \Phi w) = \lambda w \\
w_\lambda = \frac{1}{\lambda} \Phi^T \underbrace{(y - \Phi w)}_{\alpha \in \mathbb{R}^n} \\
w_\lambda = \Phi^T \alpha = \sum_{i=1}^n \alpha_i \phi(x_i)
\end{aligned}
$$

Dual:

$$
\begin{aligned}
\min_{\alpha \in \mathbb{R}^n} \|y - \Phi \Phi^T \alpha\|^2 + \lambda \alpha^T \phi \phi^T \alpha \\
(K + \lambda + I) \alpha_\lambda = y
\end{aligned}
$$

Kernal trick:

$$
\begin{aligned}
\alpha_\lambda = (K + \lambda I)^{-1} y \\
K = \Phi \Phi^T, K_{ij} = \phi(x_i)^T \phi(x_j)
\end{aligned}
$$

21.1 Representer Theorem

$$
\begin{aligned}
\lambda > 0 \\
w_\lambda = \arg\min_{w \in \mathbb{R}^D} \sum_{i=1}^n l\left(y_i w^T \phi(x_i)\right) - \lambda \|w\|^2
\end{aligned}
$$

then

$$
w_\lambda = \sum \alpha_i \phi(x_i)
$$

Classifier: for a new x,

$$
\begin{aligned}
\hat{y} = \text{sign}(\hat{w} \phi(x)) \\
= \text{sign}\left(\sum \alpha_i \phi(x_i)^T \phi(x)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \operatorname{sign}\left(\sum \alpha_i K\left(x_i, x\right)\right) \\
K\left(x, x'\right) &= \left(x^T x' + 1\right)^k \\
&\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} l\left(y_i\left(\sum_{j=1}^{n} \alpha_i\left(\phi\left(x_j\right)^T\right) \phi\left(x_i\right)\right)\right) + \lambda \sum \alpha_i \alpha_j \underbrace{\phi^T\left(x_i\right) \phi\left(x_j\right)}_{K\left(x_i, x_j\right)}
\end{aligned}
$$

Proof. WLOG, $w_\lambda \in \mathbb{R}^D$

$$
w_\lambda = \sum_{j=1}^{n} \alpha_j \phi\left(x_j\right) + u
$$

where $u^T \phi\left(x_i\right) = 0, i = 1, ..., n$

$$
\begin{aligned}
\left\|w_\lambda\right\|^2 &= \left\|\sum \alpha_i \phi\left(x_j\right) + u\right\|^2 \\
&= \left\|\sum \alpha_i \phi\left(x_j\right)\right\|^2 + \|u\|^2 \\
&= \left\|\sum \alpha_i \phi\left(x_j\right)\right\|^2
\end{aligned}
$$

□

21.2 Kernels

$$
\begin{aligned}
K\left(x, x'\right) &= \left(x^T x' + 1\right)^k \\
x \in \mathbb{R}^d, D &= \begin{bmatrix} d + K \\ K \end{bmatrix}
\end{aligned}
$$

eg. $d = 10, k = 4, D = 1001$

$$
\operatorname{sign}\left(\sum \alpha_i K\left(x_i, x\right)\right)
$$

what is this? multi-dim poly

21.3 Infinite Dim Feature Spaces

$$
\begin{aligned}
&l_p = \left\{\left(\beta_1, \beta_2, ...\right) : \sum_{i \geqslant 1} \beta_i^p < \infty\right\} \\
&p = 2, l_2
\end{aligned}
$$

Suppose we have a sequence of features $\left\{\phi_i\left(x\right)\right\}_{i \geqslant 1} \in l_2$

Define

$$
\begin{aligned}
K\left(x, x'\right) &= \sum_{i \geqslant 1} \phi_i\left(x\right) \phi_i\left(x'\right) \\
&= < \phi\left(x\right), \phi\left(x'\right) >
\end{aligned}
$$

l_2 inner product, symm

$$
\begin{aligned}
|K(x, x')|^2 &= |<\phi(x), \phi(x')>|^2 \\
&\leqslant \|\phi(x)\|^2 \|\phi(x')\|^2 \\
&< \infty
\end{aligned}
$$

Cauchy-Schwarz

21.4 Taylor Series Kernel

$$
\begin{aligned}
f(z) &= \sum_{j=0}^{\infty} a_j z^j, z \in \mathbb{R} \\
e^z &= \sum_{j=0}^{\infty} \left(\frac{z^j}{j!}\right) < \infty
\end{aligned}
$$

Exp kernel:

$$
K(x, x') = \exp\left(x^T x'\right)
$$

21.5 Gaussian Kernel

$$
\begin{aligned}
K(x, x') &= \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right) \\
&= \exp\left(-\frac{\|x\|^2}{\sigma^2}\right) \exp\left(-\frac{\|x'\|^2}{\sigma^2}\right) \exp\left(-\frac{2x^T x'}{\sigma^2}\right)
\end{aligned}
$$

21.6 Kernels

$$
\begin{aligned}
\{\phi_j(x)\}_{j \geqslant 1} \in l_2, K(x, x') =&< \phi(x), \phi(x') > \\
f(x) &= \sum_{i=1}^{N} \alpha_i K(x_i, x), x_i \in \mathbb{R}^d \\
\hat{y} &= \text{sign}(f(x)) \\
f \in \mathcal{H} &= \left\{f : f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)\right\}
\end{aligned}
$$

Hilbert Space

$$
\begin{aligned}
& f, g \in \mathcal{H} \\
& \alpha f + \beta g \in \mathcal{H} \\
& \alpha, \beta \in \mathbb{R}
\end{aligned}
$$

Lemma 3. *$\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a valid kernel iff it is symmetric and PSD for any $N \geqslant 1$ and any $x_1, ..., x_N \in \mathbb{R}^d$. Gram matrix $K_{ij} = K(x_i, x_j)$ is symm PSD.*

Proof. $x_1 ... x_N$

$$\begin{aligned}
K\left(x, x'\right) &= \phi^T\left(x\right)\phi\left(x'\right) \\
&K, N \times N \\
&v \in \mathbb{R}^N \\
v^T K v &= \sum_{i,j=1}^{N} v_i v_j \underbrace{K\left(x_i, x_j\right)}_{\phi(x_i)\phi(x_j)} = \left(\sum_i v_i \phi\left(x_i\right)\right)^T \left(\sum_j v_j \phi\left(x_j\right)\right) \\
&= \left\|\sum v_i \phi\left(x_i\right)\right\|^2 \geqslant 0
\end{aligned}$$

□

Proof. K is PSD $\Rightarrow K\left(x, x'\right)$ is an inner product
define $\phi\left(x\right)$ to be $X \rightarrow K\left(\cdot, x\right)$, "canonical feature map"

$$\mathcal{H} = \left\{f : f\left(\cdot\right) = \sum_{i=1}^{N} \alpha_i K\left(\cdot, x_i\right)\right\}, x_i \in \mathbb{R}^d$$

□

Define inner product,

$$\begin{aligned}
&< f, g >=< \sum \alpha_i K\left(\cdot, x_i\right), \sum \beta_j K\left(\cdot, x_j\right) > \\
&= \sum_{i,j=1}^{N} \alpha_i \beta_j K\left(x_i, x_j\right) \\
&= \alpha^T K \beta
\end{aligned}$$

this is valid

$$\min_{w \in \mathbb{R}^D} \sum_{i=1}^{n} l\left(y_i w^T \phi\left(x_i\right)\right) + \lambda \|w\|^2$$

What if $D > n$?

$$\begin{aligned}
K\left(x, x'\right) &= \exp\left(-\frac{\|x - k'\|^2}{\sigma^2}\right) \\
&\min_w \|y - \Phi w\|^2 + \lambda \|w\|^2 \\
\Phi &= \begin{bmatrix} \phi^T\left(x_1\right) \\ ... \\ \phi^T\left(x_n\right) \end{bmatrix}, n \times D, D >> n \\
w_\lambda &= \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T y \\
w_0 &= \lim_{\lambda \to 0} w_\lambda = \underbrace{\Phi^T}_{D \times n} \underbrace{\left(\Phi \Phi^T\right)^{-1}}_{n \times n} y
\end{aligned}$$

pseudo inverse

$$\Phi^T = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix}$$

and if $D > n$,

$$\| y - \underbrace{\Phi w_0}_{\hat{y}} \|^2 = 0$$

min norm solution

$$w_0 = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

21.7 Laplacian Kernel

$$\begin{aligned} K(x, x') &= \exp\left(-\beta \|x - x'\|\right) \\ \beta &\sim n \end{aligned}$$

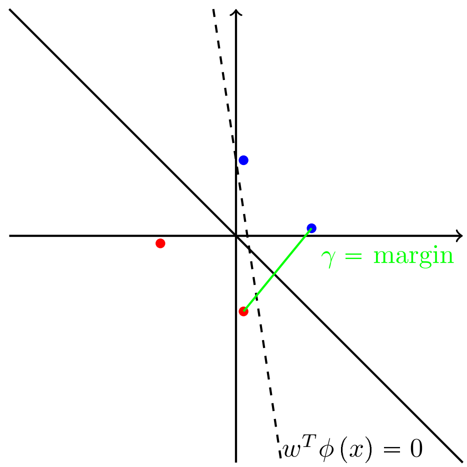
22 Lecture 22

22.1 SVMs

$$\begin{aligned} &\min_w \sum_{i=1}^{n} \left(1 - y_i w^T \phi(x_i)\right)_+ + \lambda \|w\|^2 \\ &\min_\alpha \sum_{i=1}^{n} \left(1 - y_i \sum_{j=1}^{n} \alpha_i K(x_i, x_j)\right)_+ \lambda \alpha^T K \alpha \end{aligned}$$

Dual:

$$\begin{aligned} K(x, x') &= \phi(x)^T \phi(x') \\ w^T \phi(x) &= \sum_{i=1}^{n} \alpha_i K(x_i, x) \end{aligned}$$



solid line is max margin separtor
dashed line is separator but not max margin
sum of hinge losses

$$H(w) = \sum_{i=1}^{n} \left(1 - y_i w^T \phi(x_i)\right)_+$$

linearly separable $\Rightarrow \exists w$ such that $H(w) = 0$
max-margin separator
w with smallest norm
and $y_i w^T \phi(x_i) \geqslant 1$ for all i

22.2 max-margin opt

$$\min_w \|w\|^2 \text{ such that } y_i w^T \phi(x_i) \geqslant 1$$
$$\text{such that } \sum_{i=1}^{n} \left(1 - y_i w^T \phi(x_i)\right)_+ = 0$$

Lagrangian form:

$$\min_w \sum_{i=1}^{n} \left(1 - y_i w^T \phi(x_i)\right)_+ + \underbrace{\lambda}_{\text{Lagrange mult}} \|w\|^2$$

for $\lambda > 0$ tiny

22.3 Perceptron

$$\{(x_i, y_i)\}\ \hat{y}_i = \operatorname{sign}\left(w^T x_i\right)$$
$$w_1 = \text{ init}$$
$$w_{t+1} = w_t + \mu \underbrace{\left(y_{i_t} - \hat{y}_{i_t}\right)}_{\text{error}} x_{i_t}, t \geqslant 1$$

$$y = 1, w^T x > 0$$

$$y - \hat{y} = \begin{cases} 0 & \text{if } y = 1, w^T x > 0 \\ 0 & \text{if } y = -1, w^T x < 0 \\ 2 & \text{if } y = 1, w^T x < 0 \\ -2 & \text{if } y = -1, w^T x > 0 \end{cases}$$

$$y - \hat{y} = \begin{cases} 0 & \text{if } y w^T x > 0 \\ 2y & \text{if } y w^T x < 0 \end{cases}$$

$$\begin{aligned} w_{t+1} &= w_t + \mu \left(2 y_{i_t}\right) \mathbb{1}_{\left\{y_{i_t} w_t^T x_{i_t} < 0\right\}} x_{i_t} \\ &= w_t + 2\mu \mathbb{1}_{\left\{y_{i_t} w_t^T x_{i_t} < 0\right\}} y_{i_t} x_{i_t} \end{aligned}$$

SGD with loss function l with derivative l'

$$\begin{aligned} w_{t+1} &= w_t + \gamma \left(-l' \left(y_{i_t} w_t^T x_{i_t}\right)\right) y_{i_t} x_{i_t} \\ &= w_t - \gamma \left.\frac{\partial l}{\partial w}\right|_{w = w_t} \\ \gamma &= 2\mu \end{aligned}$$

-$l'(z) = \mathbb{1}_{\{z<0\}}$
$\hat{y} = f\left(w^T x\right)$, single layer perceptron

22.4 Multilayer Neural Network

$$\hat{y} = W_L f\left(W_{L-1} \ldots \left(W_2 f\left(W_1 x + b_1\right) + b_2\right) + \ldots + b_{L-1}\right) + b_L$$

Wx affine linear map
$f(\cdot)$ nonlinear coodinate-wise
$f(v) = \begin{bmatrix} f(v_1) \\ \ldots \\ f(v_n) \end{bmatrix}, v \in \mathbb{R}^n$, "activation" function

$$\min_{(w_j, b_j)_{j=1}^L} \sum_{i=1}^n l\left(y_i \hat{y}_i\left(\{w_j, b_j\}\right)\right)$$

22.5 Two-Layer Neural Net

$$\hat{y} = W_2 f\left(W_1 + b_1\right) + b_2$$

linear in W_2, b_2, but nonlinear in w_1, b_1

22.6 Kernel Machine $= 2$ Layer Net

$$\hat{y} = \sum_{i=1}^{n} \alpha_i K\left(x_i, x\right)$$

linear in parameter

$$\begin{aligned} W_2 &= \begin{bmatrix} \alpha_1 & \ldots & \alpha_n \end{bmatrix} \\ K\left(x_i, x\right) &= f\left(x_i^T + b_i\right) \end{aligned}$$

Example 15. $\left(w_i^T x + 1\right)^k, f\left(\cdot\right) = \left(\cdot\right)^k$

Example 16. $\exp\left(x_i^T x\right), f\left(\cdot\right) = \exp\left(\cdot\right)$

Example 17. $\exp\left(-\frac{1}{2}\|x_i - x\|^2\right) = \exp\left(x_i^T x - \underbrace{\frac{1}{2}\left(\|x\|^2 + \|x_i\|^2\right)}_{b_1}\right)$

$$\begin{aligned} W_1^T &= \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \\ \hat{y} &= W_2 f\left(W_1 x + b_1\right) \\ &\begin{bmatrix} x_1^T x \\ \ldots \\ x_n^T x \end{bmatrix} \end{aligned}$$

22.7 Difference

W_1, b_1 fixed for kernel machine
layer 1 width $= n$
Kernel machines are neural nets
    Variable parameter for neural net
layer 2 width = anything
But note all neural nets are kernal mahines

23 Lecture 23

23.1 Two-Layer NNs

$$y = W_2 f\left(W_1 x + b_1\right)$$

W_1, b_1 are fixed in kernel

23.2 Convolutions NNs

1 network layer (in general)

$$y = f\left(w_i^T x + b_i\right)$$
$$W_1 = \begin{bmatrix} w_1^T \\ w_2^T \\ \dots \\ w_n^T \end{bmatrix}$$

CNN

$$y_i = f\left(\|w_i \star x\|_\infty - b_i\right)$$

convolution

$$\left(w_i \star x\right)_k = \sum_{j=1}^{d} w_{ij} x_{k-j}$$

max pooling: max output of conv

$$\|w_i \star x\|_\infty$$
$$f(z) = \max\{0, z\}$$

23.3 Backprop = SGD

23.4 Stone-Weierstrauss Thm (informal)

any continuous function $[0,1]^d$ can be approximated point-wise to arbitrary accurracy with a polynomial.

Theorem 12. *for any continuous f on $[0,1]^d$ there exists a neural net*

$$g(x) = W_2 f(W_1 x)$$
$$W_1 \in \mathbb{R}^{D\times d}$$
$$f\left(u_i^T x\right) = \left(u_i^T + 1\right)^k$$
$$W_1 = \begin{bmatrix} u_1^T \\ \dots \\ u_n^T \end{bmatrix}$$

for k and n sufficiently large, $n = D$ and

$$u_i \in \mathbb{R}^d$$
$$u_i \overset{iid}{\sim} p$$

where p is any continuous density on $[0,1]^d$

Example:

$$
\begin{aligned}
g &: \mathbb{R}^d \to \mathbb{R} \\
g(x) &= v^T f(W, x) \\
&= v^T \begin{bmatrix} \left(v_1^T x + 1\right)^k \\ \dots \\ \left(v_n^T x + 1\right)^k \end{bmatrix} \\
&= v^T \begin{bmatrix} \phi(u_1)^T \phi(x) \\ \dots \\ \phi(u_n)^T \phi(x) \end{bmatrix} \\
\phi(x) \in \mathbb{R}^D, D &= \binom{d+k}{k} \\
g(x) &= \sum_{i=1}^{n} v_i \phi(u_i)^T \phi(x) \\
&= \left(\Phi^T v\right)^T \phi(x) \\
&= w^T \phi(x) \\
\Phi^T &= \begin{bmatrix} \phi(u_1) & \dots & \phi(u_D) \end{bmatrix}
\end{aligned}
$$

general polynomial
If Φ^T is invertible

$$
v = \left(\Phi^T\right)^{-1} w
$$

Lemma 4. *If $h : \mathbb{R}^d \to \mathbb{R}$ is a polynomial function $\neq 0$ and if $u \in \mathbb{R}^d$ is a continuous random point, then $\mathbb{P}\{h(u) = 0\} = 0$*

Intuition:

$$
\phi(u_1) \in \mathbb{R}^D
$$

$v_1, \dots, v_{D-1}$ be a basis for othogonal subspace

$$
\begin{aligned}
&\phi(u_2) \\
&\mathbb{P}\left\{v_j^T \phi(u_2) = 0\right\} = 0 \\
&\Rightarrow \phi(u_2) \text{ is linearly independent of } \phi(u_1)
\end{aligned}
$$

24 Lecture 24

24.1 Probably Approx Correct Learning

$\mathcal{X}$ = feature space, $\mathcal{X} = \mathbb{R}^d, \hat{y} = f(x), f \in \mathcal{F}$
$\mathcal{Y}$ = label space, $\mathcal{Y} = \{-1, +1\}$
$\mathcal{F}$ = hypothesis space, $\mathcal{F} = \{$ linear classifiers $\}$

$l =$ loss function, $l : \mathcal{Y} \times \mathcal{Y}\mathbb{R}_+$

24.2 Empirical Risk Minimization (ERM)

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} l\left(y_i, f\left(x_i\right)\right)$$
$$\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n} \overset{iid}{\sim} P$$

Emp Risk:

$$\hat{R}\left(f\right) = \frac{1}{n} \sum_{i=1}^{n} l\left(y_i, f\left(x_i\right)\right)$$

True Risk:

$$R\left(f\right) = \mathbb{E}_{\left(x_i, y_i\right) \sim P}\left[l\left(y, f\left(x\right)\right)\right]$$
$$f^{\star} = \arg\min_{f \in \mathcal{F}} R\left(f\right)$$

Want $\hat{R}\left(f\right) \approx R\left(f\right)$ for all f, uniformly close

$$\begin{aligned}
\mathbb{P}\left\{|\hat{R}\left(f\right) - R\left(f\right)| \geqslant t\right\} &\leqslant \mathbb{P}\left\{|\hat{R}\left(f\right) - R\left(f\right)|^2 \geqslant t^2\right\} \\
&\leqslant \frac{\mathbb{E}\left[\left(\hat{R}\left(f\right) - R\left(f\right)\right)^2\right]}{t^2} \\
&= \frac{\mathbb{V}\left[\hat{R}\left(f\right)\right]}{t^2} \\
&\leqslant \frac{c^2}{4nt^2}
\end{aligned}$$

Markov's Inequality

$$\hat{R}\left(f\right) = \frac{1}{n}\left(\sum_{i=1}^{n}\left[\underbrace{l\left(y_i, f\left(x_i\right)\right]}_{\text{bounded by } c}\right)$$
$$0 \leqslant l \leqslant c$$

$l = 0$ with probability $\frac{1}{2}$
$l = c$ with porbability $\frac{1}{2}$

$$\Rightarrow \mathbb{V}\left[l\right] \leqslant \frac{c^2}{4}$$

24.3 Chernoff's Bound

$$\mathbb{P}\left\{\hat{R}\left(f\right) - R\left(f\right) \geqslant t\right\}$$

$$
\begin{aligned}
\mathbb{P}\left\{e^{\lambda\left(\hat{R}(f)-R(f)\right)} \geqslant e^{\lambda t}\right\} &\leqslant e^{-\lambda t}\mathbb{E}\left[e^{\lambda\left(\hat{R}(f)-R(f)\right)}\right] \\
\mathbb{P}\left\{\hat{R}\left(f\right)-R\left(f\right) \geqslant t\right) &\leqslant e^{-\dfrac{2nt^2}{c^2}} \\
\mathbb{P}\left\{|\hat{R}\left(f\right)-R\left(f\right)| \geqslant t\right\} &= \mathbb{P}\left\{\hat{R}\left(f\right)-R\left(f\right) \geqslant t \text{ or } R\left(f\right)-\hat{R}\left(f\right) \geqslant t\right) \\
&\leqslant 2e^{-\dfrac{2nt^2}{c^2}}
\end{aligned}
$$

by union bound

$$
\mathbb{P}\left\{\hat{R}\left(f_1\right)-R\left(f_1\right) \geqslant t \text{ or } \hat{R}\left(f_2\right)-R\left(f_2\right) \geqslant t \text{ or } ...\right\} \leqslant \text{ small bound}
$$

Assume $\mathcal{F}$ is finte.
$k = |\mathcal{F}|$ is the number of classifiers in $\mathcal{F}$.

24.4 Uniform Bound

$$
\begin{aligned}
&\mathbb{P}\left\{\hat{R}\left(f_1\right)-R\left(f_1\right) \geqslant t \text{ or } \hat{R}\left(f_2\right)-R\left(f_2\right) \geqslant t \text{ or } ... \underbrace{\hat{R}\left(f_k\right)-R\left(f_k\right) \geqslant t}_{\text{bad things}}\right\} \\
&\leqslant \sum_{i=1}^{k}\left\{|\hat{R}\left(f_i\right)-R\left(f_i\right)| \geqslant t\right\}, \text{ union bound} \\
&\leqslant \underbrace{|\mathcal{F}|2e^{-\dfrac{2nt^2}{c^2}}}_{\delta}
\end{aligned}
$$

$$
\begin{aligned}
\delta &= 2|\mathcal{F}|e^{-\dfrac{2nt^2}{c^2}} \\
\log\frac{2|\mathcal{F}|}{\delta} &= \frac{2nt^2}{c^2} \\
t &= \sqrt{\frac{c^2\log\left(\dfrac{2|\mathcal{F}|}{\delta}\right)}{2n}}
\end{aligned}
$$

with probabiliy $\geqslant 1-\delta$,

$$
\max_{f\in\mathcal{F}}|\hat{R}\left(f\right)-R\left(f\right)| \leqslant \sqrt{\frac{c^2\log\left(\dfrac{2|\mathcal{F}|}{\delta}\right)}{2n}}
$$

$$
\begin{aligned}
f^\star &= \arg\min_{f\in\mathcal{F}}\mathbb{E}\left[l\left(y, f\left(x\right)\right)\right] \\
&= \arg\min_{f\in\mathcal{F}} R\left(f\right) \\
\hat{f} &= \arg\min_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n} l\left(y_i, f\left(x_i\right)\right)
\end{aligned}
$$

$$= \arg\min_{f \in \mathcal{F}} \hat{R}(f)$$

$$\begin{aligned}
R\left(\hat{f}\right) &\leqslant \hat{R}\left(\hat{f}\right) + \sqrt{\frac{c^2 \log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{2n}} \text{ wp } \geqslant 1-\delta \\
&\leqslant \hat{R}(f^\star) + \sqrt{\frac{c^2 \log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{2n}}, \text{ since } \hat{f} \min \hat{R} \\
&\leqslant R(f^\star) + \sqrt{\frac{c^2 \log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{n}}
\end{aligned}$$

Generalization Bound

decreases like $\frac{1}{\sqrt{n}}$

inreases with $\log\left(\frac{1}{\delta}\right)$ and $\log|\mathcal{F}|$

$$\begin{aligned}
n &\geqslant \log|\mathcal{F}| \\
\Delta &= \min_{f \neq f^\star} R(f) - R(f^\star)
\end{aligned}$$

with prob $\geqslant 1-\delta$

$$\begin{aligned}
R\left(\hat{f}\right) &\leqslant R(f^\star) + \sqrt{\frac{2\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{n}} \\
\mathbb{E}\left[R\left(\hat{f}\right)\right] &\leqslant (1-\delta)\left[R(f^\star) + \sqrt{\frac{2\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{n}}\right] + \delta c, \text{ take } \delta = \frac{1}{\sqrt{n}} \\
&\leqslant R(f^\star) + \tilde{O}\left(\sqrt{\frac{\log|\mathcal{F}| + \log n}{n}}\right)
\end{aligned}$$

24.5 Infinite $\mathcal{F}$

$$\begin{aligned}
\mathcal{F} &= \left\{ \text{ all linear classifiers on } [0,1]^k \right\} \\
|\mathcal{F}_\varepsilon| &= o\left(\left(\frac{1}{\varepsilon}\right)^d\right) \\
\log|\mathcal{F}_\varepsilon| &\sim d \log\frac{1}{\varepsilon} \\
t &= \sqrt{\frac{c^2 \log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{n}}
\end{aligned}$$

24.6 Hyperparameter Tuning

$$\hat{w}_\lambda = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(1 - y_i w^T x_i\right)_+ \lambda \|w\|^2, \lambda \in \Lambda$$

want "tune" λ

$$\begin{aligned} \{(x_i, y_i)\}_{i=1}^n \\ n = n_T + n_H \\ w_\lambda = \arg\min \sum_{i=1}^{n_T} \left(1 - y_i w^T x_i\right)_+ + \lambda \|w\|^2 \end{aligned}$$

use hold out set $\{(x_i, y_i)\}_{i=n_T+1}^n$ validate, tune λ

$$\hat{R}(w_\lambda) = \frac{1}{n_H} = \sum_{i=n_T+1}^{n} \mathbb{1}_{\{y_i w_\lambda^T x_i < 0\}}$$

number of mistakes w_λ makes on hold out

$$\begin{aligned} \lambda^\star = \arg\min_{\lambda \in \Lambda} \mathbb{E}\left[\mathbb{1}_{\{y w_\lambda^T x < 0\}}\right] \\ \hat{\lambda} = \arg\min_{\lambda \in \Lambda} \hat{R}(w_\lambda) \end{aligned}$$

Assume Λ is finite

$$\begin{aligned} \Lambda = \{\lambda_1, \lambda_2, ..., \lambda_k\} \\ |\hat{R}(w_\lambda) - R(w_\lambda)| \leqslant \sqrt{\frac{2 \log \frac{2|\Lambda|}{\delta}}{n_H}}, \text{ wp } \geqslant 1 - \delta \end{aligned}$$

25 Lecture 25

Feature space $\mathcal{X}$
Label sapce $\mathcal{Y}$
Predictor $f : \mathcal{X} \to \mathcal{Y}, f \in \mathcal{F}$
Loss function $l : l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$

$$\begin{aligned} \{(x_i, y_i)\}_{i=1,2,\dots,n} \overset{iid}{\sim} P \\ R(f) = \mathbb{E}_{(x,y)\sim P}\left[l(y, f(x))\right] \\ \hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) \\ f^\star = \arg\min_{f \in \mathcal{F}} R(f) \\ \hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f) \\ \mathbb{P}\left\{|R\left(\hat{f}\right) - R(f^\star)| \geqslant t\right\} \end{aligned}$$

$$\leqslant 2|\mathcal{F}| \exp\left(-\frac{2nt^2}{c^2}\right)$$

$$t < \sqrt{\frac{\log|\mathcal{F}| + \log n}{n}}$$

25.1 Markov Ineq

$X > 0, \phi$ increasing

$$\mathbb{P}\{\phi(X) > \phi(t)\} = \mathbb{P}\{X \geqslant t\} \leqslant \frac{\mathbb{E}[X]}{t}$$

$$\mathbb{E}[X] \geqslant \mathbb{E}\left[X \mathbb{1}_{[X \geqslant t]}\right] \geqslant \mathbb{E}\left[t \mathbb{1}_{[X \geqslant t]}\right] = t\mathbb{P}\{X \geqslant t\}$$

25.2 Chebyshev Inequality

$$\begin{aligned}
&\mathbb{P}\{|X - \mathbb{E}[X]| \geqslant t\} \\
&= \mathbb{P}\left\{|X - \mathbb{E}[X]|^2 \geqslant t^2\right\} \\
&\leqslant \frac{\mathbb{V}[X]}{t^2}
\end{aligned}$$

25.3 Cherboeff Bound

$$\begin{aligned}
&\mathbb{P}\{X - \mathbb{E}[X] \geqslant t\} \\
&= \mathbb{P}\left\{\exp(s(X - \mathbb{E}[X])) \geqslant e^{st}\right\} \\
&\leqslant \frac{\mathbb{E}[\exp(s(X - \mathbb{E}[X]))]}{e^{st}} \\
&\leqslant \min_{s>0} \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st} e^{s\mathbb{E}[X]}}
\end{aligned}$$

25.4 Sub Gaussian Random Variables

Assume $\mathbb{E}[X] = 0$,

$$\mathbb{E}\left[e^{sX}\right] \leqslant e^{\frac{cs^2}{2}}, c > 0$$

$$\mathbb{P}\{|X| \geqslant t\} \leqslant \frac{e^{\frac{cs^2}{2}}}{e^{st}} = e^{\frac{cs^2}{2} - st}$$

$$\leqslant e^{-\frac{t^2}{2c}}$$

$$s = \frac{t}{c}$$

X is subGaussian with param c,

$$S_n = \sum_{i=1}^{n} x_i$$

$$\mathbb{E}\left[e^{sS_n}\right] = \mathbb{E}\left[e^{s\sum_{i=1}^{n} x_i}\right]$$
$$= \prod_{i=1}^{n} \mathbb{E}\left[e^{sX_i}\right] \leqslant \exp\left(\frac{ncs^2}{2}\right)$$
$$\mathbb{P}\left\{|S_n| \geqslant t\right\} \leqslant \exp\left(-\frac{t^2}{2nc}\right)$$
$$\mathbb{P}\left\{\frac{|1}{n}\sum x_i| \geqslant t\right\} \leqslant \exp\left(-\frac{nt^2}{2c}\right)$$

X is bounded, $x \in [a, b]$

$$\mathbb{E}\left[e^{sX}\right] \leqslant e^{\frac{(b-a)^2 s^2}{2}}$$

25.5 Shattering Coefficient of $\mathcal{F}$

$$D_n = \{(x_i, y_i)\}_{i=1,\ldots,n}$$
$$\mathcal{S}(\mathcal{F}, n) = \max_{x_1,\ldots,x_n} |\{(f(x_1), f(x_2), \ldots, f(x_n)), f \in \mathcal{F}\}|$$

Examples:
$\mathcal{F}$: 1-dim linear classifier

$$\mathcal{F}_{D_n} = \{(+,+,+), (-,-,-), (+,-,-), (-,-,+), (-,+,+), (+,+,-)\}$$
$$VC(\mathcal{F}) \geqslant 2$$

$\mathcal{F}'$: 2-dim linear classifier

$$\mathcal{F}_{D_n} = \{-1, +1\}^3$$
$$VC(\mathcal{F}) \geqslant 3$$

The total number of classifiers is,

$$2\binom{n}{d}$$

$\mathcal{F}$ shatter D_n if $\mathcal{F}_{D_n}$ contains all tuples of $\{+1, -1\}^n \Rightarrow |\mathcal{F}_{D_n}| = 2^n$.
$VC(\mathcal{F}) = k$ if k is the largest integer there exist D_k such that $\mathcal{F}$ shatter D_k
VC of d dim linear classifiers is $d + 1$

25.6 FTSLT

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| \geqslant \varepsilon\right\} \leqslant 8\mathcal{S}(\mathcal{F}, n)\, e^{-\frac{n\varepsilon^2}{32}}$$

$$= O\left(e^{0\frac{n\varepsilon^2}{c}+VC\log n}\right)$$

$$\mathcal{S}(\mathcal{F},n) \leqslant (n+1)^{VC}$$

$$D_n = \{(x_i,y_i)\}_{i=1,\dots,n}$$
$$D_n' = \{(x_i',y_i')\}_{i=1,\dots,n}$$
$$\hat{R}_n'(f) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[f(x_i')\neq y_i]}$$

Step 1:

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|\hat{R}_n(f)-R(f)|\geqslant\varepsilon\right\}\leqslant 2\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|\hat{R}_n(f)-\hat{R}_n'(f)|>\frac{\varepsilon}{2}\right\}$$

$$\{\sigma_i\}_{i=1,\dots,n} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2}\end{cases}$$

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|\hat{R}_n(f)-\hat{R}_n'(f)|>\frac{\varepsilon}{2}\right\}\\
&=\mathbb{P}\left\{\sum_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{f(x_i)\neq y_i\}}-\mathbb{1}_{\{f(x_i')\neq y_i'\}}\right)\right|\geqslant\frac{\varepsilon}{2}\right\}\\
&\leqslant\mathbb{P}\left\{\sum_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{f(x_i)\neq y_i\}}\right)\right|\geqslant\frac{\varepsilon}{4}\right\}+\mathbb{P}\left\{\sum_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{f(x_i')\neq y_i'\}}\right)\right|\geqslant\frac{\varepsilon}{4}\right\}\\
&\mathbb{P}\left\{\sum_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{f(x_i)\neq y_i\}}\right)\right|\geqslant\varepsilon\right\}\\
&=\mathbb{E}\left[\mathbb{1}_{\left\{\sup_{f\in\mathcal{F}_{D_n}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{f(x_i)\neq y_i\}}\right)\right|\geqslant\varepsilon\right\}}|D_n\right]\\
&\leqslant\mathcal{S}(\mathcal{F},n)\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\underbrace{\sigma_i\left(\mathbb{1}_{\{f(x_i)\neq y_i\}}\right)}_{iid,boundedby1,expecation0}\right|\geqslant\varepsilon|D_n\right\}\right.\\
&\leqslant\mathcal{S}(\mathcal{F},n)\,e^{-\frac{n\varepsilon^2}{32}}
\end{aligned}$$

26 Lecture 26

$$R(f) = \mathbb{P}\{f(X)\neq y\}$$

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{f(x_i) \neq y_i\}}$$
$$f \in \mathcal{F}, \hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$$
$$f^\star = \arg\min_{f \in \mathcal{F}} R(f)$$

26.1 Uniform Derivation Bound

$$\sup_{f \in \mathcal{F}} \left| R(f) - \hat{R}_n(f) \right| \leqslant 6\sqrt{\frac{\underbrace{d_{\mathcal{F}}}_{VCdim(\mathcal{F})} \log\left(\frac{n}{\delta}\right)}{n}} \text{ wp } \geqslant 1 - \delta$$

$$\begin{aligned} R\left(\hat{f}\right) &\leqslant \hat{R}\left(\hat{f}\right) + 6\sqrt{\frac{d_{\mathcal{F}} \log\left(\frac{n}{\delta}\right)}{n}} \\ &\leqslant \hat{R}(f^\star) + 6\sqrt{\frac{d_{\mathcal{F}} \log\left(\frac{n}{\delta}\right)}{n}} \\ &\leqslant R\left(\hat{f}\right) + 12\sqrt{\frac{d_{\mathcal{F}} \log\left(\frac{n}{\delta}\right)}{n}} \end{aligned}$$

generalization bound

26.2 VC Dimensions

VC dim of axis-aligned rectangles in $\mathbb{R}^d = 2d$

26.3 Sample Complexity

$$n >> d_{\mathcal{F}}$$

26.4 Neural nets

VC dim of L layer ReLU $= O(L \cdot W)$
W is total number of weights $\neq d$
L is number of layers

26.5 Cross Validation

hold out $\frac{n}{10}$

$$\hat{f} = \arg\min_{\frac{9}{10}n}$$

$$R\left(\hat{f}\right) \leqslant \hat{R}\left(\hat{f}\right) + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$$

Chernoff bound

26.6 Convex Loss Functions

$$\begin{aligned}
l(z) &= (1-z)_+ \\
\mathbb{P}\{\text{ err }\} \leqslant R(f) &= \mathbb{E}\left[\hat{l}(yf(x))\right] \\
\hat{R}(f) &= \frac{1}{n}\sum_{i=1}^{n} l(y_i f(x_i)) \\
\hat{f} &= \arg\min_{f\in\mathcal{F}} \hat{R}(f) \\
f^\star &= \arg\min_{f\in\mathcal{F}} R(f)
\end{aligned}$$

26.7 Linear Classifiers

$$\begin{aligned}
f(x) &= w^T x \\
&\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \left(1 - y_i w^T x_i\right)_+ \\
\sup_{f\in\mathcal{F}} |R(f) - \hat{R}(f)| &\leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} |R(f) - \hat{R}(f)|\right] + \sqrt{\frac{2\log\frac{1}{\delta}}{n}} \text{ wp } 1-\delta
\end{aligned}$$

McDiramid's Inequality

26.8 Rademacher Complexity

$$\begin{aligned}
\mathbb{E}\left[\sup_f |R(f) - \hat{R}(f)|\right] &\leqslant \mathbb{E}\left[\sup_f |\hat{R}'(f) - \hat{R}(f)|\right] \\
&= \mathbb{E}\left[\sup_f \frac{1}{n}|\sum_{i=1}^{n}\left(l\left(y_i' f\left(x_i'\right)\right) - l(y_i f(x_i))\right)|\right] \\
&= \mathbb{E}\left[\sup_f \frac{1}{n}\sigma_i\left(\left(l\left(y_i' f\left(x_i'\right)\right) - l(y_i f(x_i))\right)\right)|\right] \\
&\leqslant 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}|\sum_{i=1}^{n}\sigma_i l(y_i f(x_i))|\right] \\
&\sigma_i \text{ iid } \pm 1 \text{ wp } \frac{1}{2}
\end{aligned}$$

Rademacher Complexity $\mathcal{R}_n(\mathcal{F})$

$\|x_i\| \leqslant 1$ for all i

$$\begin{aligned}
\mathcal{W}_M &= \left\{w \in \mathbb{R}^d : \|w\| \leqslant M\right\} \\
&\min_{w \in \mathcal{W}_M} \sum_{i=1}^{n} \left(1 - y_i w^T x_i\right)_+ \\
\mathcal{R}_n\left(\mathcal{W}_M\right) &= 2\mathbb{E}\left[\sup_{w \in \mathcal{W}_n} \frac{1}{n} \left| \sum_{i=1}^{n} l\left(y_i w^T x_i\right) \right| \right] \\
&\leqslant 4\mathbb{E}\left[\sup_{w \in \mathcal{W}_n} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i y_i x_i \right| \right] \\
&\leqslant 4\mathbb{E}\left[\sup_{w \in \mathcal{W}_n} \frac{1}{n} \|M\| \| \sum_{i=1}^{n} \sigma_i x_i \| \right] \\
&\leqslant \frac{4M}{\sqrt{n}}
\end{aligned}$$

Theorem 13. *('17,' 18)$\mathcal{F}$ = neural nets with L layers, M, using hinge or logistic*

$$R\left(\hat{f}\right) \leqslant \hat{R}\left(\hat{f}\right) + O\left(\sqrt{\frac{LM^2}{n}}\right)$$

M is product of Frobenius norms of all weight matrix

27 Problem Set 1

27.1 Q1

No

27.2 Q2

No

27.3 Q3

Strategy 1: completely randomized,

$$\begin{aligned}
\mathbb{P}\left\{X = \hat{X}, Y = \hat{Y}\right\} &= \mathbb{P}\left\{X = \hat{X}\right\}\mathbb{P}\left\{Y = \hat{Y}\right\} \\
&= \frac{1}{6}\frac{1}{6} \\
&= \frac{1}{36}
\end{aligned}$$

Strategy 2: guess the other players number,

$$\begin{aligned}
\mathbb{P}\left\{X = Y\right\} &= \frac{1}{6} \\
&< \frac{1}{36}
\end{aligned}$$

This is minimal because,

$$\begin{aligned}
\mathbb{P}\left\{X = \hat{X}, Y = \hat{Y}\right\} &\leqslant \mathbb{P}\left\{X = \hat{X}\right\} \\
&\leqslant \frac{1}{6}
\end{aligned}$$

28 Problem Set 2

28.1 Q1

$$\begin{aligned}
\mathbb{P}\left\{g(x) \neq Y | X = x\right\} &= \mathbb{P}\left\{Y = 1 | X = x\right\}\mathbb{P}\left\{g(X) = 0 | X = x\right\} + \mathbb{P}\left\{Y = 0 | X = x\right\}\mathbb{P}\left\{g(X) = 1 | X =
\\
&= \eta(x)\left(1 - \mathbb{P}\left\{g(X) = 1 | X = x\right\}\right) + \left(1 - \eta(x)\right)\mathbb{P}\left\{g(X) = 1 | X = x\right\} \\
\mathbb{P}\left\{g(x) \neq Y | X = x\right\} - \mathbb{P}\left\{f^{\star}(x) \neq Y | X = x\right\} &= \left(2\eta(x) - 1\right)\left(\mathbb{P}\left\{f^{\star}(x) = 1 | X = x\right\} - \mathbb{P}\left\{g(x) = 1 | X = x\right\}\right) \\
&\geqslant 0
\end{aligned}$$

If $\eta(x) \geqslant \frac{1}{2}$, then,

$$\begin{aligned}
2\eta(x) - 1 \geqslant 0 \\
\mathbb{P}\left\{f^{\star}(x) = 1 | X = x\right\} - \mathbb{P}\left\{g(x) = 1 | X = x\right\} = 1 - \mathbb{P}\left\{g(x) = 1 | X = x\right\} \geqslant 0
\end{aligned}$$

If $\eta(x) < \frac{1}{2}$, then,

$$
\begin{aligned}
2\eta(x) - 1 &< 0 \\
\mathbb{P}\{f^\star(x) = 1|X = x\} - \mathbb{P}\{g(x) = 1|X = x\} = 1 - \mathbb{P}\{g(x) = 1|X = x\} &\leqslant 0
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\{g(X) \neq Y\} &= \mathbb{E}[\mathbb{P}\{g(X) \neq Y\}|X] \\
&\geqslant \mathbb{E}[\mathbb{P}\{f^\star(X) \neq Y\}|X] \\
&= \mathbb{P}\{f^\star(X) \neq Y\}
\end{aligned}
$$

28.2 Q2

$$
\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}[Z|Z < t]\mathbb{P}\{Z < t\} + \mathbb{E}[Z|Z \geqslant t]\mathbb{P}\{Z \geqslant t\} \\
&\geqslant \mathbb{E}[Z|Z \geqslant t]\mathbb{P}\{Z \geqslant t\} \\
&\geqslant t\mathbb{P}\{Z \geqslant t\}
\end{aligned}
$$

28.3 Q3

Start with showing mean = 0,

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{f(X_i)\neq Y_i} - p_f\right] &= \mathbb{P}\{f(X_i) \neq Y_i\} - p_f \\
&= 0
\end{aligned}
$$

Use Markov,

$$
\begin{aligned}
\mathbb{P}\{|\hat{p}_f - p_f| > \varepsilon\} &\leqslant \frac{\mathbb{E}\left[(\hat{p}_f - p_f)^2\right]}{\varepsilon^2} \\
&= \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq Y_i} - p_f\right)^2\right] \\
&= \frac{1}{\varepsilon^2}\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq Y_i} - p_f\right] \\
&= \frac{1}{\varepsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[\mathbb{1}_{f(X_i)\neq Y_i} - p_f\right] \\
&= \frac{1}{(n\varepsilon)^2}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{1}_{f(X_i)\neq Y_i} - 2p_f\mathbb{1}_{f(X_i)\neq Y_i} + p_f^2\right] \\
&= \frac{1}{(n\varepsilon)^2}\sum_{i=1}^{n}\left(p_f - p_f^2\right) \\
&= \frac{p_f(1 - p_f)}{n\varepsilon^2}
\end{aligned}
$$

28.4 Q4

By convexity,

$$\begin{aligned}g(x) &\geqslant g(t) + g'(t)(x-t) \\ \mathbb{E}[g(x)] &\geqslant g(t) + g'(t)(\mathbb{E}[x]-t)\end{aligned}$$

With $t = \mathbb{E}[x]$ and $g(x) = x^2$

$$\mathbb{E}\left[x^2\right] \geqslant \mathbb{E}[x]^2$$

28.5 Q5

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda\end{aligned}$$

28.6 Q6

$$\begin{aligned}\mathbb{P}\{X_1 + X_2 = n\} &= \sum_{i=0}^{n} \mathbb{P}\{X_1 = i, X_2 = n-i\} \\ &= \sum_{m=0}^{n} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{n-i}}{(n-i)!} \\ &= e^{-\lambda_1-\lambda_2} \frac{1}{n!} \sum_{i=0}^{n} \binom{n}{i} \lambda_1^i \lambda_2^{n-i} \\ &= e^{-\lambda_1-\lambda_2} \frac{(\lambda_1+\lambda_2)^n}{n!} \\ \mathbb{P}\{X_1 = k | X_1 + X_2 = n\} &= \frac{\mathbb{P}\{X_1 = k, X_1 + X_2 = n\}}{\mathbb{P}\{X_1 + X_2 = n\}} \\ &= \frac{e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!}}{e^{-\lambda_1-\lambda_2} \frac{(\lambda_1+\lambda_2)^n}{n!}} \\ &= \binom{n}{k} \delta^k (1-\delta)^{n-k}, \delta = \frac{\lambda_1}{\lambda_1+\lambda_2}\end{aligned}$$

28.7 Q7

Define $\eta(x,k) = \mathbb{P}\{Y = k | X = x\}$ for $k = 1, ..., m$, then

$$f^\star(x) = \arg\max_k \eta(x,k)$$

28.8 Q8

The minimum error is $p_f = 1 - \mathbb{E}_x\left[\max_k \eta\left(x, k\right)\right]$

28.9 Q9

An estimator is,

$$\hat{p}_f = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)\neq Y_i}$$

28.10 Q10

Use the result from Q3,

$$\begin{aligned}
&\mathbb{P}\left\{p_f - \hat{p}_f > \varepsilon\right\} \leqslant \frac{p_f\left(1-p_f\right)}{n\varepsilon^2} \leqslant \frac{0.25}{n\varepsilon^2}\\
&\Rightarrow \mathbb{P}\left\{p_f > 0.05 + 0.05\right\} \leqslant \frac{0.25}{1000 \cdot 0.05^2}\\
&\Rightarrow p_f < 0.1
\end{aligned}$$

29 Problem Set 3

29.1 Q1

$$\begin{aligned}
\mathbb{P}\left\{f\left(x\right) \neq Y | X = x\right\} - \mathbb{P}\left\{f^\star\left(x\right) \neq Y | X = x\right\} &= \eta\left(x\right)\left(1 - \mathbb{1}_{f(x)=1}\right) + \left(1-\eta\left(x\right)\right)\mathbb{1}_{f(x)=1} - \eta\left(x\right)\left(1 - \mathbb{1}_{f^\star(x)=1}\right) - \left(1-\eta\left(x\right)\right)\mathbb{1}\\
&= \eta\left(x\right)\left(2\mathbb{1}_{f^\star(x)=1} - 2\mathbb{1}_{f(x)=1}\right) + \mathbb{1}_{f(x)=1} - \mathbb{1}_{f^\star(x)=1}\\
&= \left(2\eta\left(x\right) - 1\right)\left(\mathbb{1}_{f^\star(x)=1} - \mathbb{1}_{f(x)=1}\right)
\end{aligned}$$

This is,

$$\begin{cases}
0 & \text{if } \eta\left(x\right) \geqslant \frac{1}{2} \text{ and } \tilde{\eta}\left(x\right) \geqslant \frac{1}{2}\\
0 & \text{if } \eta\left(x\right) < \frac{1}{2} \text{ and } \tilde{\eta}\left(x\right) < \frac{1}{2}\\
2\left(\eta\left(x\right) - \frac{1}{2}\right) \leqslant 2\left(\eta\left(x\right) - \tilde{\eta}\left(x\right)\right) & \text{if } \eta\left(x\right) \geqslant \frac{1}{2} \text{ and } \tilde{\eta}\left(x\right) < \frac{1}{2}\\
2\left(\eta\left(x\right) - \frac{1}{2}\right) \leqslant 2\left(\eta\left(x\right) - \tilde{\eta}\left(x\right)\right) & \text{if } \eta\left(x\right) < \frac{1}{2} \text{ and } \tilde{\eta}\left(x\right) \geqslant \frac{1}{2}
\end{cases}$$

Therefore,

$$\mathbb{P}\left\{f\left(x\right) \neq Y | X = x\right\} - \mathbb{P}\left\{f^\star\left(x\right) \neq Y | X = x\right\} \leqslant 2|\eta\left(x\right) - \tilde{\eta}\left(x\right)|$$

29.2 Q2

1. The loss function is,

$$\begin{cases}
c_{01} & \text{if } f\left(x\right) = 0, y = 1\\
c_{10} & \text{if } f\left(x\right) = 1, y = 0
\end{cases}$$

   Then the expect loss given f and $f^\star$ is,

$$\mathbb{E}\left[l\left(f, X, Y\right)|X = x\right] - \mathbb{E}\left[l\left(f^{\star}, X, Y\right)|X = x\right] = c_{01}\eta\left(x\right)\left(1 - \mathbb{1}_{f(x)=1}\right) + c_{10}\left(1 - \eta\left(x\right)\right)\mathbb{1}_{f(x)=1} - c_{01}\eta\left(x\right)\left(1 - \mathbb{1}_{f^{\star}(x)=1}\right.$$

$$= \eta\left(x\right)\left(-\left(c_{01} + c_{10}\right)\mathbb{1}_{f(x)=1} + \left(c_{01} + c_{10}\right)\mathbb{1}_{f^{\star}(x)=1}\right) + c_{10}\left(\mathbb{1}_{f(x)=1} - \mathbb{1}_{f^{\star}(x)=1}\right))$$
$$= \left(\eta\left(x\right)\left(c_{01} + c_{10}\right) - c_{10}\right)\left(\mathbb{1}_{f^{\star}(x)=1}\right) - \mathbb{1}_{f(x)=1}))$$

To minimize the expected loss,
Set $f^{\star}\left(x\right) = 1$ if and only if,

$$\eta\left(x\right)\left(c_{01} + c_{10}\right) - c_{10} \geqslant 0$$
$$\eta\left(x\right) \geqslant \frac{c_{10}}{c_{01} + c_{10}}$$

Therefore, the optimal classifier is,

$$\begin{cases} 1 & \text{if } \eta\left(x\right) \geqslant \dfrac{c_{10}}{c_{01} + c_{10}} \\ 0 & \text{otherwise} \end{cases}$$

2. No. $f^{\star}\left(x\right)$ does minimize the average probability of error for any p.

1. The same as (a), π_i is already contained in $\eta\left(x\right) = \dfrac{p\left(x|Y = 1\right)\pi_1}{p\left(x\right)}$.

29.3 Q3

1. The optimal classifier is $\hat{y} = x_1, R^{\star} = 0$

1. All possible data points are $y = 1$ and $x = (1, 1)$ or $(1, -1)$ and $y = -1$ and $x = (-1, 1)$ or (-1, $-1)$, the probability of error is, assuming the original data are $(1, x_1)$ and (-1, x_2), and the new data $(1, \delta)$ has label 1,

$$\mathbb{P}\left\{f_n\left(x\right) \neq 1\right\} = \frac{1}{2}\mathbb{P}\left\{\delta \neq x_1\right\}\mathbb{P}\left\{x_1 \neq x_2\right\}$$
$$= \frac{1}{2}\frac{1}{2}\frac{1}{2}$$
$$= \frac{1}{8}$$

2. Assume the new data has label 1, and the training data are $u = (1, u_2, ..., u_d)$ and $v = (-1, v_2, ..., v_d)$, Let $U = \sum_j |u_j - x_j| \sim \text{ Bin }\left(d - 1, \frac{1}{2}\right)$ and $V = \sum_j |v_j - x_j| - 1 \sim \text{ Bin }\left(d - 1, \frac{1}{2}\right)$,
and note that $d - 1 + U - V \sim \text{ Bin }\left(2d - 2, \frac{1}{2}\right)$,

$$\mathbb{P}\left\{f_n\left(x\right) \neq 1\right\} = \mathbb{P}\left\{\|u - x\| < \|v - x\|\right\} + \frac{1}{2}\mathbb{P}\left\{\|u - x\| = \|v - x\|\right\}$$
$$= \mathbb{P}\left\{U + 1 < V\right\} + \frac{1}{2}\mathbb{P}\left\{1 + U = V\right\}$$
$$= \mathbb{P}\left\{d - 1 + V - U > d\right\} + \frac{1}{2}\mathbb{P}\left\{d - 1 + V - U = d\right\}$$

$$= \sum_{i=d+1}^{2d-2} \binom{2d-2}{i} \frac{1}{2^{2d-2}} + \frac{1}{2} \binom{2d-2}{d} \frac{1}{2^{2d-2}}$$

3. As $d \to \infty$, Bin $\left(2d-2, \frac{1}{2}\right)$ is approximately $N\left(d-1, \frac{d-1}{2}\right)$ or $N\left(d, \frac{d}{2}\right)$

$$\begin{aligned}
\mathbb{P}\left\{f_n(x) \neq 1\right\} &= \mathbb{P}\left\{d-1+V-U > d\right\} + \frac{1}{2}\mathbb{P}\left\{d-1+V-U = d\right\} \\
&= \frac{1}{2} + 0 \\
&= \frac{1}{2}
\end{aligned}$$

29.4 Q4

1. The MLE is,

$$\begin{aligned}
\hat{y}(x) &= \arg\max_l p(y=l|x) \\
&= \arg\max_l p(x|y=l)\, p(y=l) \\
&= \arg\max_l \log p(x|y=1) + \log p(y=l) \\
&= \arg\max_l \frac{-1}{2} \log|\Sigma_l| - \frac{1}{2}(x-\mu_l)^T \Sigma_{l^{-1}} (x-\mu_l) + \log \pi_l
\end{aligned}$$

   Given data,

$$\hat{y}(x) = \arg\max_l \frac{-1}{2} \log|\hat{\Sigma}_l| - \frac{1}{2}(x-\hat{\mu}_l)^T \hat{\Sigma}_{l^{-1}} (x-\hat{\mu}_l) + \log \hat{\pi}_l$$

2. Use a common covariance matrix $\hat{\Sigma}$ instead of individual Σ_l,

$$\begin{aligned}
\hat{y}(x) &= \arg\max_l \frac{-1}{2} \log|\hat{\Sigma}| - \frac{1}{2}(x-\hat{\mu}_l)^T \left(\hat{\Sigma}\right)^{-1} (x-\hat{\mu}_l) + \log \hat{\pi}_l \\
&= \arg\max_l -\frac{1}{2}(x-\hat{\mu}_l)^T \hat{\Sigma}^{-1} (x-\hat{\mu}_l) + \log \hat{\pi}_l \\
&= \arg\max_l -x^T \left(\hat{\Sigma}\right)^{-1} x + 2\hat{\mu}_l^T \hat{\Sigma}^{-1} x - \hat{\mu}_l^T \hat{\Sigma}^{-1} \hat{\mu}_l + \log \hat{\pi}_l \\
&= \arg\max_l -x^T \left(\hat{\Sigma}\right)^{-1} x + 2\hat{\mu}_l^T \hat{\Sigma}^{-1} x - \hat{\mu}_l^T \hat{\Sigma}^{-1} \hat{\mu}_l + \log \hat{\pi}_l \\
&= \arg\max_l \left(2\hat{\mu}_l^T \hat{\Sigma}^{-1}\right) x - \left(\hat{\mu}_l^T \hat{\Sigma}^{-1} \hat{\mu}_l + \log \hat{\pi}_l\right)
\end{aligned}$$

29.5 Q5

Use Markov's Inequality,

$$\mathbb{P}\left\{X \geqslant t\right\} \leqslant \mathbb{P}\left\{X^2 \geqslant t^2\right\} \leqslant \frac{\mathbb{E}\left[X^2\right]}{t^2}$$

29.6 Q6

The event $\{X+Y>t\} \subseteq \left\{X > \frac{t}{2}\right\} \cup \left\{Y > \frac{t}{2}\right\}$

$$\mathbb{P}\{X+Y>t\} \leqslant \mathbb{P}\left\{X > \frac{t}{2}\right\} + \mathbb{P}\left\{Y > \frac{t}{2}\right\}$$

29.7 Q7

1. The optimal Bayes classifier is always,

$$f^{\star}(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

To simplify $\eta(x)$,

$$\begin{aligned}
\eta(x) > \frac{1}{2} &\Leftrightarrow \mathbb{P}\{Y=1|X=x\} > \frac{1}{2} \\
&\Leftrightarrow \mathbb{P}\{X=x|Y=1\}\mathbb{P}\{Y=1\} > \frac{1}{2}\mathbb{P}\{X=x\} \\
&\Leftrightarrow 2\mathbb{P}\{X=x|Y=1\}\mathbb{P}\{Y=1\} > \mathbb{P}\{X=x|Y=1\}\mathbb{P}\{Y=1\} + \mathbb{P}\{X=x|Y=-1\}\mathbb{P}\{Y=-1\} \\
&\Leftrightarrow \mathbb{P}\{X=x|Y=1\}\mathbb{P}\{Y=1\} > \mathbb{P}\{X=x|Y=-1\}\mathbb{P}\{Y=-1\} \\
&\Leftrightarrow \log\mathbb{P}\{X=x|Y=1\} + \log\mathbb{P}\{Y=1\} > \log\mathbb{P}\{X=x|Y=-1\} + \log\mathbb{P}\{Y=-1\} \\
&\Leftrightarrow -\frac{1}{2}\frac{1}{\sigma}(x-\theta)^T(x-\theta) + \log\pi_1 > -\frac{1}{2}\frac{1}{\sigma}(x+\theta)^T(x+\theta) + \log\pi_{-1} \\
&\Leftrightarrow -\frac{1}{2}\frac{1}{\sigma}\left(x^Tx - 2x^T\theta + \theta^T\theta - x^Tx - 2x^T\theta - \theta^T\theta\right) > \log\pi_{-1} - \log_{\pi_1} \\
&\Leftrightarrow \frac{2}{\sigma}x^T\theta > \log\pi_{-1} - \log\pi_1 \\
&\Leftrightarrow x^T\theta > \frac{\sigma}{2}(\log\pi_{-1} - \log\pi_1)
\end{aligned}$$

Therefore, the optimal Bayes classifier is,

$$f^{\star}(x) = \begin{cases} 1 & \text{if } x^T\theta > \frac{\sigma}{2}(\log\pi_{-1} - \log\pi_1) \\ -1 & \text{otherwise} \end{cases}$$

when $\pi_1 = \pi_{-1}$,

$$f^{\star}(x) = \begin{cases} 1 & \text{if } x^T\theta > 0 \\ -1 & \text{otherwise} \end{cases}$$

2. The MLE is,

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta}\sum_{i=1}^{n}\log\mathbb{P}\{x_i, y_i|\theta\} \\
&= \arg\max_{\theta}\sum_{i=1}^{n}\frac{-1}{2}\frac{1}{\sigma}(x_i - \theta y_i)^T(x_i - \theta y_i) + \log\pi_{y_i}
\end{aligned}$$

$$
\begin{aligned}
&= \arg\min_\theta \sum_{i=1}^n (x_i - \theta y_i)^T (x_i - \theta y_i) \\
&= \arg\min_\theta \sum_{i=1}^n \left(x_i^T x_i - 2y_i\theta^T x_i + y_i^2\theta^T\theta\right) \\
&= \arg\min_\theta \sum_{i=1}^n \left(-2y_i\theta^T x_i + \theta^T\theta\right) \\
&= \theta : \sum_{i=1}^n -2y_i x_i + 2\theta = 0 \\
&= \frac{1}{n}\sum_{i=1}^n x_i y_i
\end{aligned}
$$

3. The plug-in classifier is

$$
\hat{f}(x) = \begin{cases} 1x^T\hat{\theta} > \dfrac{\hat{\sigma}}{2}\left(\log\hat{\pi}_{-1} - \log\hat{\pi}_1\right) & \\ -1 & \text{otherwise} \end{cases}
$$

where,

$$
\begin{aligned}
\hat{\sigma} &= \frac{1}{nd}(x_i - \theta y_i)^T (x_i - \theta y_i) \\
\hat{\pi}_j &= \frac{n_{y_i=j}}{n}
\end{aligned}
$$

4. The error probability for x with true label -1 is,

$$
\mathbb{P}\left\{\tilde{f}(x) = 1\right\} = \mathbb{P}\left\{x^T\hat{\theta} > \frac{\hat{\sigma}}{2}\left(\log\hat{\pi}_{-1} - \log\hat{\pi}_1\right)\right\}
$$

where

$$
\begin{aligned}
x &\sim N\left(-\theta, \sigma^2 I\right) \\
\theta &\sim N\left(\theta, \frac{\sigma^2}{n} I\right)
\end{aligned}
$$

5. The error probability from the previous part, and define $e_1 = x + \theta$ and $e_2 = \hat{\theta} - \theta$

$$
\begin{aligned}
\mathbb{P}\left\{\tilde{f}(x) = 1\right\} &\leqslant \mathbb{P}\left\{x^T\hat{\theta} > 0\right\} \\
&= \mathbb{P}\left\{(e_1 - \theta)^T (e_2 + \theta) > 0\right\} \\
&= \mathbb{P}\left\{\left(e_1^T e_2 + e_1^T\theta - e_2^T\theta - \theta^T\theta\right) > 0\right\} \\
&\leqslant \mathbb{P}\left\{\left(e_1^T - e_2^T\right)\theta > \frac{1}{2}\theta^T\theta\right\} + \mathbb{P}\left\{e_1^T e_2 > \frac{1}{2}\theta^T\theta\right\} \\
&\leqslant \frac{\mathbb{E}\left[\left(\left(e_1^T - e_2^T\right)\theta\right)^2\right]}{\left(\frac{1}{2}\theta^T\theta\right)^2} + \frac{\mathbb{E}\left[\left(e_1^T e_2\right)^2\right]}{\left(\frac{1}{2}\theta^T\theta\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\sigma^2\left(1+\frac{1}{n}\right)\theta^T\theta+\sigma^4\frac{d^2}{n}}{\left(\frac{1}{2}\theta^T\theta\right)^2} \\
&= O\left(\max\left\{\frac{\sigma^2}{\theta^T\theta}, \frac{\sigma^4 d^2}{n\left(\theta^T\theta\right)^2}\right\}\right)
\end{aligned}
$$

To reduce the error, both of the following needs to hold,

$$
\begin{aligned}
\sigma^2 &<< \theta^T\theta \\
\sigma^4 &<< \frac{n}{d^2}\left(\theta^T\theta\right)^2
\end{aligned}
$$

or,

$$
\sigma^2 << \min\left\{\theta^T\theta, \frac{\sqrt{n}}{d}\left(\theta^T\theta\right)\right\}
$$

30 Problem Set 4

30.1 Q1

1. Use formula,

$$
\begin{aligned}
x \sim N(\mu, \Sigma) &\Rightarrow Ax+b \sim N\left(A\mu+b, A\Sigma A^T\right) \\
w^T x &\sim N\left(0, w^T\Sigma w\right)
\end{aligned}
$$

2. Similarly,

$$
\begin{aligned}
z &= Ax \sim N\left(0, A\Sigma A^T\right) \\
&\Rightarrow A\Sigma A^T = I \\
&\Rightarrow A = \Sigma^{-\frac{1}{2}} = UD^{-\frac{1}{2}}U^T
\end{aligned}
$$

30.2 Q2

1. $p(x_i|\theta) = \frac{1}{\theta}\mathbb{1}_{x\leqslant\theta}$

$$
\begin{aligned}
\hat{\theta}_n &= \arg\max_\theta p(x_1, ..., x_n|\theta) \\
&= \arg\max_\theta \prod_{i=1}^n p(x_i|\theta) \\
&= \arg\max_\theta \frac{1}{\theta^n}\mathbb{1}_{\max\limits_i (x_i)} \leqslant \theta \\
&= \max_i \{x_i\}
\end{aligned}
$$

2. The CDF,

$$
\begin{aligned}
F_{\hat{\theta}_n}(x) &= \mathbb{P}\left\{\hat{\theta}_n \leqslant x\right\} \\
&= \mathbb{P}\left\{\max_i \{x_i\} \leqslant x\right\} \\
&= \prod_{i=1}^{n} \mathbb{P}\{x_i \leqslant x\} \\
&= \left(\frac{x}{\theta}\right)^n
\end{aligned}
$$

Then the PDF,

$$
\begin{aligned}
f_{\hat{\theta}_n}(x) &= F'_{\hat{\theta}_n}(x) \\
&= n\frac{x^{n-1}}{\theta^n}
\end{aligned}
$$

3. The MSE is,

$$
\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] &= \int_0^\theta (x-\theta)^2\, n\frac{x^{n-1}}{\theta^n} dx \\
&= n\int_0^\theta \frac{x^{n+1}}{\theta^n} - 2\frac{x^n}{\theta^{n-1}} + \frac{x^{n-1}}{\theta^{n-2}} dx \\
&= \frac{n}{\theta^n}\int_0^\theta x^{n+1} - 2x^n\theta + x^{n-1}\theta^2 dx \\
&= \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2} - 2\frac{x^{n+1}}{n+1}\theta + \frac{x^n}{n}\theta^2\right]_{x=0}^{\theta} \\
&= \frac{n}{\theta^n}\left(\frac{\theta^{n+2}}{n+2} - 2\frac{\theta^{n+2}}{n+1} + \frac{\theta^{n+2}}{n}\right) \\
&= n\theta^2\frac{(n+1)\,n - 2\,(n+2)\,n + (n+1)\,(n+2)}{n\,(n+1)\,(n+2)} \\
&= \frac{2\theta^2}{(n+1)\,(n+2)} \\
&\to 0 \text{ as } n \to \infty
\end{aligned}
$$

30.3 Q3

1. The first derivative test for e_1 is,

$$
\begin{aligned}
&\sum_{i=1}^{n} \operatorname{sign}\left(\hat{\theta}_1 - x_i\right) = 0 \\
&\Rightarrow \hat{\theta}_1 = \text{ median } (x)
\end{aligned}
$$

The first derivative test for e_2 is,

$$
\sum_{i=1}^{n} 2\left(\hat{\theta}_2 - x_i\right) = 0
$$

$$\Rightarrow \hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. The data is $(0.9, 1.0, 1.1)$,

$$\hat{\theta}_1 = 1$$
$$\hat{\theta}_2 = 1$$

3. The data is $(0.9, 1.1, 100)$,

$$\hat{\theta}_1 = \frac{102}{3}$$
$$\hat{\theta}_2 = 1.1$$

30.4 Q4

Use the invariance property and find the MLE for θ first,

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \prod_{i=1}^{N} p\left(\tau_i | \theta\right) \\
&= \arg\max_{\theta} \sum_{i=1}^{N} \left(\log\theta - \tau_i \theta\right) \\
&= \theta : \frac{N}{\theta} - \sum_{i=1}^{N} \tau_i = 0 \\
&= \hat{\theta} = \frac{N}{\sum_{i=1}^{N} \tau_i}
\end{aligned}$$

Then,

$$\begin{aligned}
\hat{\mathbb{P}}\left\{\tau < 10\right\} &= \int_0^{10} \hat{\theta} e^{-\tau\hat{\theta}} d\tau \\
&= 1 - e^{-10\hat{\theta}} \\
&= 1 - e^{-10 \frac{N}{\sum_{i=1}^{N} \tau_i}}
\end{aligned}$$

31 Problem Set 5

31.1 Q1

Use Factorization Theorem,

$$p\left(x | a, b\right) = \prod_{i=1}^{N} \frac{1}{b-a} \mathbb{1}_{a \leqslant x_i \leqslant b}$$

$$
\begin{aligned}
&= \frac{1}{(b-a)^N} \mathbb{1}_{a \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n \leqslant b} \\
&= \frac{1}{(b-a)^N} \mathbb{1}_{a \leqslant x_{(1)} \leqslant x_{(n)} \leqslant b}
\end{aligned}
$$

Therefore, $x_{(1)}$ and $x_{(n)}$ are sufficient

31.2 Q2

1. The covariance matrix is,

$$
\begin{aligned}
\Sigma &= \mathbb{E}\left[XX^T\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[X\right]^T \\
&= Q - \mu^2 ee^T
\end{aligned}
$$

Use Sherman-Morrison formula,

$$
\Sigma^{-1} = Q^{-1} + \frac{\mu^2 Q^{-1} ee^T Q^{-1}}{1 - \mu^2 e^T Q^{-1} e}
$$

Then,

$$
\begin{aligned}
\log p\left(x|\mu\right) &= \left(X - \mu e\right)^T \Sigma^{-1} \left(X - \mu e\right) \\
&= X^T \Sigma^{-1} X + \mu^2 e^T \Sigma^{-1} e - 2\mu X^T \Sigma^{-1} e \\
&= X^T Q^{-1} X + \frac{\mu^2 X^T Q^{-1} ee^T Q^{-1} X}{1 - \mu^2 e^T Q^{-1} e} - 2\mu \left(X^T Q^{-1} e + \frac{\mu^2 X^T Q^{-1} ee^T Q^{-1} e}{1 - \mu^2 e^T Q^{-1} e}\right)
\end{aligned}
$$

Therefore, by Factorization Theorem,

$$
\begin{aligned}
t\left(X\right) &= X^T Q^{-1} e \\
&= X^T \left(\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \frac{1}{\rho + 1} \left(X_1 + X_2\right)
\end{aligned}
$$

Note that, $X_1 + X_2$ is another sufficient statistics.

2. $\mathbb{E}\left[X_1\right] = \mu$

1. First note that,

$$
\mathbb{E}\left[X_1 | T_X = X_1 + X_2\right] = \frac{T_X}{2}
$$

Then,

$$
\mathbb{E}\left[\frac{T_X}{2}\right] = \mu
$$

and,

$$\begin{aligned}
\mathbb{V}\left[\frac{T_X}{2}\right] &= \frac{1}{4}\left(\mathbb{V}\left[X_1\right] + \mathbb{V}\left[X_2\right] + 2Cov\left[X_1, X_2\right]\right) \\
&= \frac{1}{2}\left(1+\rho\right) \\
&\leqslant \frac{1}{2}\left(1+1\right) \\
&= 1 \\
&= \mathbb{V}\left[X_1\right]
\end{aligned}$$

31.3 Q3

1. MLE is invariant,

$$\begin{aligned}
\hat{\phi} &= \hat{p}_1 - \hat{p}_2 \\
&= \frac{x_1}{n_1} - \frac{x_2}{n_2}
\end{aligned}$$

2. The Log-Like is for p_i is $x_i \log p_i + (n_i - x_i)\log\left(1-p_i\right)$

$$\begin{aligned}
\frac{\partial \log\left(x|p\right)}{\partial p_i} &= \frac{x_i}{p_i} - \frac{n_i - x_i}{1-p_i} \\
\frac{\partial^2 \log\left(x|p\right)}{\partial p_i^2} &= -\frac{x_i}{p_i^2} - \frac{n_i - x_i}{1-p_i}^2 \\
\mathbb{E}\left[x_i\right] &= n_i p_i \\
\mathbb{E}\left[-\frac{\partial^2 \log\left(x|p\right)}{\partial p_i^2}\right] &= \frac{n_i}{p_i} - \frac{n_i}{1-p_i} \\
&= \frac{n_i}{p_i\left(1-p_i\right)} \\
\mathbb{E}\left[-\frac{\partial \log\left(x|p\right)}{\partial p_i \partial p_j}\right] &= 0
\end{aligned}$$

Therefore,

$$I\left(p\right) = \begin{bmatrix} \frac{n_1}{p_1\left(1-p_1\right)} & 0 \\ 0 & \frac{n_2}{p_2\left(1-p_2\right)} \end{bmatrix}$$

3. Use MLE Asypt theorem,

$$\begin{aligned}
&\hat{p} \sim N\left(p, I\right) \\
&\hat{\phi} \sim N\left(p_1 - p_2, \frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}\right)
\end{aligned}$$

4. Use results from (c),

$$\begin{aligned}\mathbb{P}\left\{|\hat{\phi}-\phi|>0.01\right\}<0.05 \Rightarrow \mathbb{P}\left\{\frac{|\hat{\phi}-\phi|}{\sigma}>\frac{0.01}{\sigma}\right\}<0.05\\ \Rightarrow \frac{0.01}{\frac{1}{\sqrt{2n}}}>2\\ \Rightarrow n>20000\end{aligned}$$

5. NO

31.4 Q4

1. For the estimator $\hat{N}_1=\frac{2}{n}\left(\sum_{i=1}^{n}x_i\right)-1$,

$$\begin{aligned}\mathbb{E}\left[x_i\right]&=\sum_{i=1}^{N}\frac{i}{N}\\ &=\frac{N+1}{2}\\ \mathbb{E}\left[x_i^2\right]&\sum_{i=1}^{N}\frac{i^2}{N}\\ &=\frac{(N+1)(2N+1)}{6}\\ \mathbb{E}\left[\hat{N}_1\right]&=\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i\right]-1\\ &=\frac{2}{n}n\frac{N+1}{2}\\ &=N\\ \mathbb{V}\left[\hat{N}_1\right]&=\frac{4}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[x_i\right]\\ &=\frac{4}{n^2}n\left(\frac{(N+1)(2N+1)}{6}-\frac{(N+1)^2}{4}\right)\\ &=\frac{N^2-1}{3n}\\ \text{MSE}\left[\hat{N}_1\right]&=\frac{N^2-1}{3n}\end{aligned}$$

$\hat{N}_1$ is unbiased.

2. MLE,

$$\begin{aligned}\hat{N}_2&=\arg\max_{N}p\left(x|N\right)\\ &=\arg\max_{N}\sum_{i=1}^{n}-\log\left(N\right)+\log\left(\mathbb{1}_{x_i\leqslant N}\right)\end{aligned}$$

$$
\begin{aligned}
&= \arg\max_N -n\log(N) + \log\left(\mathbb{1}_{x_{(n)}\leqslant N}\right) \\
&= x_{(n)} \\
\mathbb{E}\left[\hat{N}_2\right] &= \sum_{i=1}^{N} i\mathbb{P}\left\{\hat{N}_2 = i\right\} \\
&= \sum_{i=1}^{N} i\left(\left(\frac{i}{N}\right)^N - \left(\frac{i-1}{N}\right)^N\right) \\
&= \frac{1}{N^N}\left(\sum_{i=1}^{N} ii^N - \sum_{i=1}^{N-1}(i+1)\,i^N\right) \\
&= \frac{1}{N^N}\left(N^{N+1} + \sum_{i=1}^{N-1} ii^N - \sum_{i=1}^{N-1}(i+1)\,i^N\right) \\
&= N - \frac{1}{N^N}\sum_{i=1}^{N-1} i^N \\
\mathbb{E}\left[\hat{N}_2^2\right] &= \sum_{i=1}^{N} i^2\mathbb{P}\left\{\hat{N}_2 = i\right\} \\
&= \sum_{i=1}^{N} i^2\left(\left(\frac{i}{N}\right)^N - \left(\frac{i-1}{N}\right)^N\right) \\
&= \frac{1}{N^N}\left(\sum_{i=1}^{N} i^2 i^N - \sum_{i=1}^{N-1}(i+1)^2\,i^N\right. \\
&= N^2 - \frac{1}{N^N}\sum_{i=1}^{N}(2i+1)\,i^N \\
\mathbb{V}\left[\hat{N}_2^2\right] &= N^2 - \frac{1}{N^N}\sum_{i=1}^{N}(2i+1)\,i^N - \left(N - \frac{1}{N^N}\sum_{i=1}^{N-1} i^N\right)^2 \\
\text{MSE}\left[\hat{N}_2\right] &= N^2 - 2N\mathbb{E}\left[\hat{N}_2\right] + \mathbb{E}\left[\hat{N}_2^2\right] \\
&= N^2 - 2N\left(N - \frac{1}{N^N}\sum_{i=1}^{N-1} i^N\right) + N^2 - \frac{1}{N^N}\sum_{i=1}^{N}(2i+1)\,i^N \\
&= \frac{2}{N}\sum_{i=1}^{N-1} i^N - \frac{1}{N^N}\sum_{i=1}^{N}(2i+1)\,i^N
\end{aligned}
$$

3. MLE is biased. Use Rao-Blackwell, and define,

$$\hat{N}_3 = \mathbb{E}\left[\hat{N}_1|x_{(n)}\right]$$

31.5 Q5

No

31.6 Q6

No

32 Sample Midterm

32.1 Q1

1. Guess and check,

$$w^{\star} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

2. Use formula,

$$\begin{aligned}
\hat{w} &= \left(X^T X\right)^{-1} X^T y \\
&= \begin{bmatrix} \frac{4}{3} \\ \frac{4}{3} \end{bmatrix} \\
\hat{y} &= \begin{bmatrix} \frac{4}{3} \\ -\frac{2}{3} \\ -\frac{2}{3} \end{bmatrix} \\
\hat{y} &= \operatorname{sign}\left(x^T \hat{w}\right)
\end{aligned}$$

32.2 Q2

1. Least Squares

$$\min_w \|y - Xw\|^2$$

2. A lot

$$\hat{w} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \hat{w} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

3. 0

1. No

1. 3

32.3 Q3

1. Sample means for each $y_i = j \in \{000, 001, 010, 011, ..., 111\}$.

$$\hat{\mu}_j = \frac{1}{\#_{y_i = j}} \sum_{i : y_i = j} x_i$$

$$\hat{\Sigma}_j = \frac{1}{\#_{y_i=j}} \sum_{i:y_i=j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T$$

2. Naive Bayes

$$x : 000 = \arg\max_j p(x|y=j)$$

3. Histogram

$$p(y=j) = \frac{\#_{y_i=j}}{n}$$

4. Marginal

$$p(x) = \sum_j p(x|y=j)\, p(y=j)$$

5. Marginal

$$\begin{aligned} p(x|y_i=1) &= \frac{p(x, y_i=1)}{p(y_i=1)} \\ &= \frac{\sum_{y_i=1} p(x|y)\, p(y)}{\sum_{y_i=1} p(y)} \end{aligned}$$

6. Naive Bayes: has disease if $p(x|y=000) < p(x|y \neq 000)$, where,

$$p(x|y \neq 0) = \frac{\sum_{y \neq 000} p(x|y)\, p(y)}{\sum_{y \neq 000} p(y)}$$

32.4 Q4

1. Algorithm 5

1. Chernoff bound for mean is,

$$\pm\sqrt{\frac{\log\left(\frac{10}{\delta}\right)}{2n}}$$

For $\delta = 0.05$ and $n = 50$, this is ± 0.23
The standard deviation is,

$$\sigma = n^{-\frac{1}{2}}$$

For $n = 50$, this is ± 0.14

2. No

32.5 Q5

1. Let the classes by ± 1 $p\left(y=1|x\right) = \dfrac{\exp\left(w_1^T x\right)}{\exp\left(w_1^T x\right) + \exp\left(w_{-1}^T x\right)})$
   $= \dfrac{1}{1+\exp\left(\left(w_{-1}-w_1\right)^T x\right)})$
   Define $w = w_{-1} - w_1$

2. Given generative model,

$$\begin{aligned}
p\left(y=k|x\right) &= \frac{p\left(x|y=k\right)p\left(y=k\right)}{p\left(x\right)} \\
&= \frac{e^{-\frac{1}{2}\left(x-\theta_k\right)^T\left(x-\theta_k\right)}\frac{1}{c}}{\sum_{j=1}^{c} e^{-\frac{1}{2}\left(x-\theta_j\right)^T\left(x-\theta_k\right)}\frac{1}{c}} \\
&= \frac{e^{-\theta_k^T x + \theta_k^T\theta_k}}{\sum_{j=1}^{c} e^{-\theta_j^T x + \theta_j^T\theta_j}} \\
&= \left(\frac{e^{-\theta_k^T x}}{\sum_{j=1}^{c} e^{-\theta_j^T x}}\right.
\end{aligned}$$

   since $\|\theta_k\| = \|\theta_j\|$ for each j.

3. The gradient is,

$$\begin{aligned}
\frac{\partial}{\partial\theta_k} - \log\left(p\left(y|x\right)\right) &= \frac{\partial}{\partial\theta_k}\left(-\theta_y^T x + \log\left(\sum_{j=1}^{c} e^{\theta_j^T x}\right)\right) \\
&= -\mathbb{1}_{y=k}x + \frac{e^{-\theta_k^T x}}{\sum_{j=1}^{c} e^{-\theta_j^T x}}x \\
&= \left(-\mathbb{1}_{y=k} + p\left(y=k|x\right)\right)x
\end{aligned}$$

32.6 Q6

1. Uniform sampling without replacement,

$$\mathbb{E}\left[\sum_{j\in S_i}\exp\left(w_j^T x_i\right)\right] = \frac{m-1}{c-1}\sum_{j\neq i}\exp\left(w_j^T x_i\right)$$

2. Use Popoviciu, and -1 $\leqslant w_j^T x_i \leqslant 1$

$$\mathbb{V}\left[\sum_{j\in S_i} \exp\left(w_j^T x_i\right)\right] \leqslant \frac{1}{4}(m-1)\left(e - \frac{1}{e}\right)^2$$

3. Use Markov,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{j\in S_i} \exp\left(w_j^T x_i\right) - \mu\right| \geqslant \varepsilon\right\} \\
&= \mathbb{P}\left\{\left(\sum_{j\in S_i} \exp\left(w_j^T x_i\right) - \mu\right)^2 \geqslant \varepsilon^2\right\} \\
&\leqslant \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\sum_{j\in S_i} \exp\left(w_j^T x_i\right) - \mu\right)^2\right] \\
&\leqslant \frac{1}{\varepsilon^2}\sigma^2
\end{aligned}$$

Use $\delta = \frac{1}{\varepsilon^2}\sigma^2$,

$$\varepsilon = \sqrt{\frac{\sigma^2}{\delta}}$$

Therefore,

$$\mathbb{P}\left\{\sum_{j\in S_i} \exp\left(w_j^T x_i\right) - \mu \leqslant \sqrt{\frac{\sigma^2}{\delta}}\right\} = 1 - \delta$$

4. With probabiliy $1 - \delta$,

$$\sum_{j\in S_i} \exp\left(w_j^T x_i\right) \leqslant \frac{m-1}{c-1} \sum_{j\neq i} \exp\left(w_j^T x_i\right) + \sqrt{\frac{\sigma^2}{\delta}}$$

Therefore,

$$\begin{aligned}
&p\left(y_i = k|x_i\right) - \tilde{p}\left(y_i = k|x_i\right) \\
&\leqslant \frac{\exp\left(w_k^T x_i\right)}{\sum_{j=1}^{c} \exp\left(w_j^T x_i\right)} + \frac{\exp\left(w_k^T x_i\right)}{\exp\left(w_k^T x_i\right) + \frac{m-1}{c-1}\sum_{j\neq i} \exp\left(w_j^T x_i\right) + \sqrt{\frac{\sigma^2}{\delta}}}
\end{aligned}$$

32.7 Q7

1. $X \sim \text{Binomial}\left(n, \frac{\pi}{4}\right)$, estimate by,

$$\hat{\pi} = \frac{4X}{n}$$

2. Mean and variance are,

$$\begin{aligned}
\mathbb{E}[\hat{\pi}] &= \mathbb{E}\left[\frac{4X}{n}\right] \\
&= \frac{4}{n} n \frac{\pi}{4} \\
&= \pi \\
\mathbb{V}[\hat{\pi}] &= \left(\frac{4}{n}\right)^2 n \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right) \\
&= \frac{1}{n}\pi(4 - \pi)
\end{aligned}$$

3. Use Chebyshev,

$$\begin{aligned}
\mathbb{P}\{|\hat{\pi} - \pi| \geqslant 0.001\} \leqslant \frac{\mathbb{V}[\hat{\pi}]}{0.001^2} \\
= \frac{1}{n}\pi(4 - \pi)\, 1000000
\end{aligned}$$

32.8 Q8

1. Factorization Theorem,

$$\begin{aligned}
p(x_1, x_2|\mu) &= c(\mu, \theta) \exp\left(-\frac{1}{2}\frac{(x_1 - \mu)^2}{\sigma_1^2} + \frac{(x_2 - \mu)^2}{\sigma_2^2}\right) \\
&= c(\mu, \theta) \exp\left(\mu\left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}\right)\right) \\
\Rightarrow T(x_1, x_2) &= \frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}
\end{aligned}$$

2. The expectation,

$$\mathbb{E}[T(x_1, x_2)] = \mu\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

3. Unbiased estimator,

$$f(x_1, x_2) = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} T(x_1, x_2)$$

4. Same

32.9 Q9

1. MLE is,

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{n} \log(\alpha) - (1 + \alpha)\log(x_i)$$

$$
\begin{aligned}
&= \alpha : \frac{n}{\alpha} - \sum_{i=1}^{n} \log(x_i) = 0 \\
&= \frac{n}{\sum_{i=1}^{n} \log(x_i)}
\end{aligned}
$$

2. A sufficient statistic is,

$$T(x) = \sum_{i=1}^{n} \log(x_i)$$

3. Yes.

$$\hat{\alpha} = \frac{n}{T(x)}$$

32.10 Q10

$$
\begin{aligned}
D(p\|q) &= \mathop{\mathbb{E}}_{p}\left[\frac{\log(p(x))}{\log(q(x))}\right] \\
&= \mathop{\mathbb{E}}_{\theta}\left[\frac{1}{2\sigma^2}\left((x-\theta^\star)^2-(x-\theta)^2\right)\right] \\
&= \mathop{\mathbb{E}}_{\theta}\left[\frac{1}{2\sigma^2}\left(2x\theta-2x\theta^\star-\theta^2+(\theta^\star)^2\right)\right] \\
&= \mathop{\mathbb{E}}_{\theta}\left[\frac{1}{2\sigma^2}\left(2\theta^2-2\theta\theta^\star-\theta^2+(\theta^\star)^2\right)\right] \\
&= \frac{1}{2\sigma^2}\left(\theta-\theta^\star\right)^2 \\
\frac{\partial^2 D(\theta\|\theta^\star)}{\partial\theta^2} &= \frac{1}{\sigma^2} \\
\mathbb{E}\left[-\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\right] &= \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\left(\frac{1}{2\sigma^2}(x-\theta)^2\right)\right] \\
&= \mathbb{E}\left[-\left(-\frac{1}{\sigma^2}\right)\right] \\
&= \frac{1}{\sigma^2}
\end{aligned}
$$

32.11 Q11

1. Poisson distribution,

$$
\begin{aligned}
p(y) &= \frac{1}{k!}\lambda^y e^{-\lambda} \\
&= \frac{1}{k!}\exp\left(y\log(\lambda)-\lambda\right) \\
&= b(y)\exp\left(\theta y - a(\theta)\right), \theta = \log(\lambda)
\end{aligned}
$$

2. $\theta = \log(\lambda)$

1. The loglikelihood,

$$l(x|\theta) = \sum_{i=1}^{n} \left(y_i w^T x_i - \exp\left(w^T x_i\right) - \log(y_i!)\right)$$

2. The derivative,

$$\frac{\partial}{\partial w} l(x|w) = \sum_{i=1}^{n} \left(y_i x_i - \exp\left(w^T x_i\right) x_i\right)$$

32.12 Q12

1. No

1. No

1. Second is incorrect.

1. Use bounds,

$$\gamma < \frac{2}{\lambda_{\max}\left(X^T X\right)}$$

where,

$$\begin{aligned}(\lambda - 5)(\lambda - 6) - 9 &= 0\\ \lambda^2 - 11\lambda + 21 &= 0\\ \lambda &= \frac{11}{2} \pm \frac{1}{2}\sqrt{37} < \frac{17}{2}\end{aligned}$$

Therefore,

$$\begin{aligned}\gamma &= 0.2\\ &< \frac{2}{\frac{17}{2}} = \frac{4}{17}\end{aligned}$$

2. Use formula,

$$\begin{aligned}w_1 &= w_0 + \gamma x^T \left(y - x^T w\right)\\ &= \begin{bmatrix} -1 \\ 0.8 \\ -0.7 \end{bmatrix}\end{aligned}$$

33 Midterm

33.1 Q1

1. No.

1. No.

1. One point is misclassified.

1. Empirical error rate is,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{y}\neq y}$$

Actual error rate is,

$$p = \mathbb{P}\left\{\hat{y} \neq y\right\}$$

Expected value and variance of empirical error rate is,

$$\begin{aligned}
\mathbb{E}\left[\hat{p}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{y}\neq y}\right] \\
&= \frac{1}{16}\sum_{i=1}^{16} p \\
&= p \\
\mathbb{V}\left[\hat{p}\right] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{y}\neq y}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\left[\mathbb{1}_{\hat{y}\neq y}\right] \\
&= \frac{1}{n^2} np\left(1-p\right) \\
&= \frac{p\left(1-p\right)}{n}
\end{aligned}$$

Use Markov Inequality (Chebyshev), with $n = 16$ in the diagram,

$$\begin{aligned}
\mathbb{P}\left\{p - \hat{p} \geqslant \frac{1}{4}\right\} \leqslant \mathbb{P}&\left\{\left|\hat{p} - \mathbb{E}\left[\hat{p}\right]\right| \geqslant \frac{1}{16}\right\} \\
&= \mathbb{V}\left[\hat{p}\right]\cdot 16 \\
&= 16\frac{p\cdot(1-p)}{n} \\
&\leqslant \frac{16}{4n} \\
&= \frac{1}{4}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left\{p \leqslant \hat{p} + \frac{1}{4}\right\} &= 1 - \mathbb{P}\left\{p - \hat{p} \geqslant \frac{1}{4}\right\} \\
&\geqslant \frac{3}{4}
\end{aligned}$$

33.2 Q2

1. No

1. Same as usual.

1. Same as usual.

1. Same.

33.3 Q3

1. The densities are,

$$p\left(x|y=0\right) = 1\mathbb{1}_{x\in[0,1]}$$
$$p\left(x|y=1\right) = \frac{1}{\theta}\mathbb{1}_{x\in[0,\theta]}$$

   Therefore, the optimal classifier is,

$$\hat{y} = \begin{cases} 1 & \text{if } x \leqslant \theta \\ 0 & \text{if } x > \theta \end{cases}$$

   The probability of error is,

$$\begin{aligned}\mathbb{P}\left\{\hat{y} \neq y\right\} &= \mathbb{P}\left\{y=0, x \leqslant \theta\right\} + \mathbb{P}\left\{y=1, x > \theta\right\} \\ &= \mathbb{P}\left\{x \leqslant \theta | y=0\right\} \cdot \mathbb{P}\left\{y=0\right\} + \mathbb{P}\left\{x > \theta | y=1\right\} \cdot \mathbb{P}\left\{y=1\right\} \\ &= \theta \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} \\ &= \frac{\theta}{2}\end{aligned}$$

2. The usual MLE,

$$\begin{aligned}\hat{\theta} &= \max_{\theta} \prod_{i=1}^{n} \left(\frac{1}{\theta}\mathbb{1}_{x\in[0,\theta]}\right) \\ &= \max_{\theta} \left(\frac{1}{\theta^n}\mathbb{1}_{\max\limits_i x_i \in [0,\theta]}\right) \\ &= \max_{i} x_i\end{aligned}$$

3. The probability is,

$$\begin{aligned}\mathbb{P}\left\{\hat{\theta} \leqslant (1-\varepsilon)\,\theta\right\} &= \mathbb{P}\left\{\max_i x_i \leqslant (1-\varepsilon)\,\theta\right\} \\ &= \left(\mathbb{P}\left\{x_i \leqslant (1-\varepsilon)\,\theta\right\}\right)^n \\ &= \left(\frac{(1-\varepsilon)\,\theta}{\theta}\right)^n \\ &= (1-\varepsilon)^n\end{aligned}$$

4. The probability is given $\hat{\theta}$,

$$\begin{aligned}
\mathbb{P}\{\hat{y} \neq y\} &= \mathbb{P}\left\{y = 0, x \leqslant \hat{\theta}\right\} + \mathbb{P}\{y = 1, x > \theta\} \\
&= \mathbb{P}\left\{x \leqslant \hat{\theta} | y = 0\right\} \cdot \mathbb{P}\{y = 0\} + \mathbb{P}\left\{x > \hat{\theta} | y = 1\right\} \cdot \mathbb{P}\{y = 1\} \\
&= \hat{\theta} \cdot \frac{1}{2} + \frac{\theta - \hat{\theta}}{\theta} \cdot \frac{1}{2}
\end{aligned}$$

Then the probability that the plug in classifier is at most $\frac{\varepsilon}{2}$ greater than the minimum is,

$$\begin{aligned}
\mathbb{P}\left\{\mathbb{P}\{\hat{y} \neq y\} \leqslant \mathbb{P}\{\hat{y} \neq y\} + \frac{\varepsilon}{2}\right\} &= \mathbb{P}\left\{\hat{\theta} \cdot \frac{1}{2} + \frac{\theta - \hat{\theta}}{\theta} \cdot \frac{1}{2} \leqslant \frac{\theta}{2} + \frac{\varepsilon}{2}\right\} \\
&= \mathbb{P}\left\{\theta\hat{\theta} + \theta - \hat{\theta} - \theta^2 - \varepsilon\theta \leqslant 0\right\} \\
&= \mathbb{P}\left\{\theta(1 - \varepsilon) - \hat{\theta} \leqslant \theta\left(\theta - \hat{\theta}\right)\right\} \\
&\geqslant \mathbb{P}\left\{\theta(1 - \varepsilon) - \hat{\theta} \leqslant 0\right\} \\
&= \mathbb{P}\left\{\hat{\theta} \geqslant (1 - \varepsilon)\theta\right\}
\end{aligned}$$

33.4 Q4

1. Use as test set,

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{n} \mathbb{1}_{\hat{y}_i \neq y_i}$$

2. Only use data in that class,

$$\hat{p}_j = \frac{\sum_{i=1}^{n} \mathbb{1}_{\hat{y}_i \neq y_i, y_i = j}}{\sum_{i=1}^{n} \mathbb{1}_{y_i = j}}$$

3. Use the binomial probability again,

$$\begin{aligned}
\mathbb{V}[\hat{p}] &= \frac{p(1 - p)}{m} \\
\mathbb{V}[\hat{p}_j] &= \frac{p_j(1 - p_j)}{n_j}
\end{aligned}$$

4. Use the one with smallest test set error, use the one with smallest,

$$k^\star = \arg\min_k \left(\max_j \hat{p}_{j,k}\right)$$

Can also use the confidence interval with Chernoff bounds,

$$[L_{j,k}, U_{j,k}] = \left[\hat{p}_{j,k} \pm \sqrt{\frac{\log\left(\frac{6l}{\delta}\right)}{2n_j}}\right]$$

For worse case best performance, use the one with smallest

$$k^\star = \arg\min_k \left(\max_j \hat{p}_{j,k} + \sqrt{\frac{\log\left(\frac{6l}{\delta}\right)}{2n_j}}\right)$$

For a fixed δ, confident that $k^\star$ is the best if,

$$\hat{p}_{j,k^\star} + \sqrt{\frac{\log\left(\frac{6l}{\delta}\right)}{2n_j}} \geqslant \hat{p}_{j,k} - \sqrt{\frac{\log\left(\frac{6l}{\delta}\right)}{2n_j}} \ \forall\ k \neq k^\star$$

34 Problem Set 6

34.1 Q1

1. Find the mean and the mode,

$$\begin{aligned}
\arg\max_\theta p(\theta;\alpha) &= \arg\max_\theta \theta^{\alpha-1}(1-\theta)^{\alpha-1} \\
&= \arg\max_\theta (\alpha-1)\log(\theta) + (\alpha-1)\log(1-\theta) \\
&= \theta : \frac{1}{\theta} - \frac{1}{1-\theta} = 0 \\
&= \theta : 2\theta = 1 \\
&= \frac{1}{2}
\end{aligned}$$

$$\mathbb{E}[X] = \int_\theta \mathrm{Beta}(\alpha,\alpha)\,\theta^\alpha(1-\theta)^{\alpha-1}$$
$$= \int_\theta \mathrm{Beta}(\alpha+1,\alpha)\,\theta^\alpha(1-\theta)^{\alpha-1}$$

$$\begin{aligned}
&= \frac{\Gamma(\alpha+1)}{\Gamma(2\alpha+1)}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)} \\
&= \frac{\alpha}{2\alpha} \\
&= \frac{1}{2}
\end{aligned}$$

2. The posterior is,

$$p(\alpha|\theta) \propto p(\theta|\alpha)\,p(\alpha)$$

$$= \text{Beta}(\alpha, \alpha)\, \theta^{\alpha-1} (1-\theta)^{\alpha-1} \theta^{s} (1-\theta)^{N-s}$$

$$\propto \theta^{\alpha-1+s} (1-\theta)^{\alpha-1+N-s}$$

This is the kernal of another Beta distribution.

3. The mean is,

$$\frac{\alpha+s}{2\alpha+n}$$

As $n \to \infty, s \to n\theta$

$$\frac{\alpha+s}{2\alpha+n} \to \theta$$

34.2 Q2

From Q1,

$$\hat{\theta} = \frac{\alpha+x}{2\alpha+n}$$

Since $x \sim \text{Bin}(\theta, n)$,

$$\begin{aligned} M\left(\hat{\theta}\right) &= \mathbb{V}\left[\hat{\theta}\right] + \left(\theta - \mathbb{E}\left[\hat{\theta}\right]\right)^2 \\ &= \frac{n\theta(1-\theta)}{(2\alpha+n)^2} + \left(\theta - \frac{\alpha+n\theta}{2\alpha+n}\right)^2 \\ &= \frac{n\theta(1-\theta) + \alpha^2(2\theta-1)^2}{(2\alpha+n)^2} \end{aligned}$$

Need the derivative to be 0,

$$\begin{aligned} & n(1-2\theta) + 4\alpha^2(2\theta-1) = 0 \\ & \Rightarrow 4\alpha^2 = n \\ & \Rightarrow \alpha = \frac{1}{2}\sqrt{n} \end{aligned}$$

Therefore, the minimax optimal estimator is,

$$\hat{\theta} = \frac{\sqrt{n}+2x}{2\sqrt{n}+2n}$$

34.3 Q3

No.

34.4 Q4

Maximize each coordinate separately,

$$
\begin{aligned}
&\min_{w_i} (y_i - w_i)^2 + \lambda \mathbb{1}_{w_i \neq 0} \\
&\Rightarrow -2(y_i - w_i) + \lambda = 0 \text{ or } w_i = 0 \\
&\Rightarrow w_i = y_i \mathbb{1}_{y_i^2 > \lambda}
\end{aligned}
$$

34.5 Q5

1. SGD step is,

$$w_{t+1} = w_t + \gamma_t x_i \left(y_i - x_i^T w_t\right)$$

2. SGD step is,

$$w_{t+1} = w_t + \gamma_t x_i \left(y_i - x_i^T w_t\right) - \lambda \gamma_t w_t$$

3. SGD step is,

$$w_{t+1} = w_t + \gamma_t x_i \left(y_i - x_i^T w_t\right) - \text{sign}\left(w_t\right) \gamma_t$$

4. SGD step is,

$$w_{t+1} = w_t + \gamma_t \frac{-y_i x_i \exp\left(-y_i w^T x_i\right)}{1 + \exp\left(-y_i w^T x_i\right)}$$

5. SGD step is,

$$w_{t+1} = w_t + \gamma_t y_i x_i \mathbb{1}_{y_i w_t^T x_i > 1}$$

6. SGD step is,

$$w_{t+1} = w_t + \gamma_t y_i x_i \mathbb{1}_{y_i w_t^T x_i > 1} - \lambda \gamma_t w_t$$

35 Problem Set 7

35.1 Q1

1. Show that $v^T K v \geqslant 0$ for all v,

$$
\begin{aligned}
v^T K v &= v^T x x^T v \\
&= \left(v^T x\right)^2 \\
&\geqslant 0
\end{aligned}
$$

2. Use sums and constant products preserve PSD,

$$\left(1 + x^T x'\right)^p = \sum_{i=1}^{n} \binom{p}{i} \left(x^T x'\right)^i$$

3. Same as (a)

35.2 Q2

1. Show that $v^T K v \geqslant 0$ for all v,

$$\begin{aligned} v^T K v &= v^T \left(K_1 + K_2\right) v \\ &= v^T K_1 v + v^T K_2 v \\ &\geqslant 0 \end{aligned}$$

2. Use eigenvalue decomposition,

$$\begin{aligned} K_1 &= \sum_{i=1}^{n} \alpha_i a_i a_i^T \\ K_2 &= \sum_{i=1}^{n} \beta_i b_i b_i^T \\ v^T K v &= v^T \left(K_1 \cdot K_2\right) v \\ &= v^T \left(\sum_{i=1}^{n} \alpha_i a_i a_i^T \cdot \sum_{i=1}^{n} \beta_i b_i b_i^T\right) v \\ &= v^T \left(\sum_{ij} \alpha_i \beta_j \left(a_i \cdot b_j\right) \left(a_i \cdot b_j\right)^T\right) v \\ &= \sum_{ij} \alpha_i \beta_j \left(v^T a_i \cdot b_j\right)^2 \\ &\geqslant 0 \end{aligned}$$

3. From (a) and (b)

1. Taylor expansion and use (c)

35.3 Q3

1. If $f(x) = 1$,

$$\begin{aligned} & \|\phi(x) - \mu_1\| \leqslant \|\phi(x) - \mu_0\| \\ & \Rightarrow \|\phi(x)\|^2 - 2 < \mu_1, \phi(x) > + \|\mu_1\|^2 \leqslant \|\phi(x)\|^2 - 2 < \mu_0, \phi(x) > + \|\mu_0\|^2 \\ & \Rightarrow < \mu_1 - \mu_0, \phi(x) > \geqslant \frac{1}{2}\left(\|\mu_1\|^2 + \|\mu_0\|^2\right) \\ & \Rightarrow < w, \phi(x) > + b \geqslant 0 \end{aligned}$$

2. The first term is,

$$
\begin{aligned}
&< w, \phi(x) >=< \mu_1, \phi(x) > - < \mu_0, \phi(x) > \\
&=< \frac{1}{n_1} \sum_{y_i=1} \phi(x_i), \phi(x) > - < \frac{1}{n_0} \sum_{y_i=0} \phi(x_i), \phi(x) > \\
&= \sum_{y} \frac{1}{n_y} \sum_{y_i=y} < \phi(x_i), \phi(x) > \\
&= \sum_{y} \frac{1}{n_y} \sum_{y_i=y} k(x_i, x)
\end{aligned}
$$

3. Use representer theorem,

$$
\begin{aligned}
L(\alpha) &= \sum_{i=1}^{n} \left(y_i - < \sum_{j=1}^{n} \alpha_j \phi(x_j), \phi(x_i) > \right)^2 + \lambda < \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{i=1}^{n} \alpha_i \phi(x_i) > \\
&= \sum_{i=1}^{n} \left(\left(y_i - \sum_{j=1}^{n} \alpha_j K_{ij} \right)^2 + \lambda \sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} \right) \\
\frac{\partial L}{\partial \alpha_j} &= \sum_{i=1}^{n} \left(2K_{ij} \left(y_i - \sum_{j=1}^{n} \alpha_j K_{ij} \right) + 2\lambda \alpha_i K_{ij} \right) \\
&= \sum_{i=1}^{n} 2K_{ij} \left(y_i - \sum_{k=1}^{n} \alpha_k K_{ik} \right) + \sum_{i=1}^{n} 2\lambda \alpha_i K_{ij}
\end{aligned}
$$

With only row i_t from the data,

$$
\begin{aligned}
\frac{\partial L}{\partial \alpha_j} &= 2K_{i_t j} \left(y_{i_t} - \sum_{k=1}^{n} \alpha_k K_{i_t k} \right) + 2\lambda \alpha_{i_t} K_{i_t j} \\
&= 2K_{i_t j} \left(y_{i_t} - K_{i_t} \alpha \right) + 2\lambda \alpha_{i_t} K_{i_t j}
\end{aligned}
$$

Put the derivatives together for all j to get the gradient,

$$
\nabla L = 2K_{i_t}^T \left(y_{i_t} - K_{i_t} \alpha \right) + 2\lambda K_{i_t} \alpha_{i_t}
$$

Note that α and α_t are the same thing, and add this to the SGD step,

$$
\begin{aligned}
\alpha_{t+1} &= \alpha_t - \gamma_t \nabla L \\
&= \alpha_t - \gamma_t \left(2K_{i_t}^T \left(y_{i_t} - K_{i_t} \alpha \right) + 2\lambda K_{i_t} \alpha_{i_t} \right. \\
&= \left(1 - 2\lambda \gamma_t K_{i_t} e_{i_t}^T \right) \alpha_t - 2\gamma_t K_{i_t}^T \left(K_{i_t} \alpha_t - y_{i_t} \right)
\end{aligned}
$$

where $e_{i_t}^T$ is $\left[0, 0, ..., 0, \underbrace{1}_{\text{position } i_t}, 0, ..., 0 \right]$.
Therefore, the SGD step is,

$$
\alpha_{t+1} = \alpha_t - \gamma_t \left(2K_{i_t}^T \left(K_{i_t} \alpha - y_{i_t} \right) + 2\lambda K_{i_t} \alpha_t \right)
$$

35.4 Q4

1. There are d terms in the form x_{ij}, d terms in the form x_{ij}^2 and $\binom{d}{2}$ terms in the form $x_{ij}x_{ik}$ and 1 constant term,

$$1 + 2d + \binom{d}{2} = \frac{1}{2}(d+1)(d+2)$$

   Therefore the rank is at most,

$$\min\left\{n, \frac{1}{2}(d+1)(d+2)\right\}$$

2. The rank is at most,

$$\min\left\{n-1, \frac{1}{2}(d+1)(d+2)\right\}$$

35.5 Q5

All three are quadratic,
Left: $\text{sign}(x_1, x_2)$
Middle: $\text{sign}(ax_1 + b)$
Right: $\text{sign}\left(a(x_1 - h)^2 + b(x_1 - k)^2 + r\right)$

36 Problem Set 8

36.1 Q1

1. Larger than 0 parts.

$$\begin{aligned} y_1 &= \max\{x_1, 0\} \\ y_2 &= \max\{x_2, 0\} \end{aligned}$$

2. The required mapping is,

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ &= a_{11}\max\{0, x_1\} - a_{11}\max\{0, -x_1\} + a_{12}\max\{0, x_2\} - a_{12}\max\{0, -x_2\} \end{aligned}$$

   Similar for the other one.

3. The specific mapping is,

$$\begin{aligned} y_1 &= \frac{1}{2}(x_1 - x_2) \\ y_2 &= \frac{1}{2}(x_2 - x_1) \\ &= -y_1 \end{aligned}$$

4. The region is,

$$
\begin{aligned}
y_1 &\geqslant y_2 \geqslant 0 \\
y_1 &= \max\{x_1, 0\} + \max\{x_2, 0\} \\
y_2 &= \max\{x_1, 0\}
\end{aligned}
$$

36.2 Q2

1. A tree can be trained to have zero error on any training set, just split on every data point.

1. Error occurs if the curve pass through a cell: 11 cells out of 64 on both diagrams.

$$\frac{11}{64}$$

2. Suppose there are m cells, and $\frac{1}{2}$ of cells are on one side, then for both histogram and decision trees,

$$\mathbb{P}\{y = 1 | x_i \in B_i, y_i = 0\} = p\frac{1}{2} + (1-p)\frac{1}{2} = \frac{1}{2}$$

3. The question is asking for the number of nodes in a tree with $2\sqrt{m}$ leaves. The fewest is a balanced tree,

$$4\sqrt{m}$$

4. Same as (d)

$$2O\left(m^{d-1}\right) = O\left(m^{d-1}\right)$$

36.3 Q3

The collection of predictors are rankings,

$$\mathcal{F} = \{f_h : h \text{ is a ranking }\}$$

The size is $m! = O\left(m^m\right)$,

$$
\begin{aligned}
\log(m!) &= O(m \log m) \\
\mathbb{E}\left[R\left(\hat{f}\right)\right] &\leqslant R(f^\star) + O\left(\sqrt{\frac{\log|\mathcal{F}| + \log n}{n}}\right) \\
&= R(f^\star) + \sqrt{\frac{m \log m + \log n}{n}}
\end{aligned}
$$

36.4 Q4

Hinge loss ignores the error in the positive half, the classifier is the mid point between the separation boundary, 5.11 and 6.1. Therefore, the classifier is,

$$\text{sign}\left(x - \frac{1}{2}(5.11 + 6.1)\right)$$

The support vector is the minimizer of the dual,

$$\alpha = (0, 0, 1, 1)$$

36.5 Q5

1. y is constant

$$\hat{y} = x_1$$

2. Solve the following system

$$\begin{aligned} w_1 + w_2 &= 1 \\ w_1 + w_3 &= 1 \\ w &= (1 + c, -c, -c)\,. \end{aligned}$$

3. Minimize by choosing c,

$$\begin{aligned} \min_w \|w\|^2 &= \min_c (1+c)^2 + c^2 + c^2 \\ &= \min_c 3c^2 + 2c + 1 \\ &\Rightarrow 6c + 2 \\ &\Rightarrow c = -\frac{1}{3} \\ &\Rightarrow w = \left(\frac{2}{3}, -\frac{1}{2}, -\frac{1}{2}\right) \end{aligned}$$

4. The minimization is,

$$\begin{aligned} \min_\alpha \|y - XX^T\alpha\|^2 &= \min_\alpha \| \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \|^2 \\ &= \min_\alpha \| \begin{bmatrix} 1 - 2\alpha_1 - \alpha_2 \\ 1 - \alpha_1 - 2\alpha_2 \end{bmatrix} \|^2 \\ &= \min_\alpha (1 - 2\alpha_1 - \alpha_2)^2 + (1 - \alpha_1 - 2\alpha_2)^2 \\ &\Rightarrow 1 - 2\alpha_1 - \alpha_2 = 1 - \alpha_1 - 2\alpha_2 = 0 \\ &\Rightarrow \alpha_1 = \alpha_2 = \frac{1}{3} \end{aligned}$$

$$\Rightarrow w = X^T \alpha = \left(\frac{2}{3}, -\frac{1}{2}, -\frac{1}{2}\right)$$

5. Use L_1 norm,

$$\min_c |1+c| + |-c| + |-c| \Rightarrow w = (1,0,0)$$

36.6 Q6

1. Consider the linear model

$$\begin{aligned} y|w &\sim N\left(Xw, \sigma_v^2 I\right), w \sim N\left(0, \sigma_w^2 I\right) \\ \hat{w} &= \arg\max_w \log p(y|w) + \log p(w) \\ &= \arg\min_w \frac{1}{2\sigma_v^2}\|y - Xw\|^2 + \frac{1}{2\sigma_w^2}\|w\|^2 \\ &= \arg\min_w \|y - Xw\|^2 + \lambda\|w\|^2, \lambda = \frac{\sigma_v^2}{\sigma_w^2} \end{aligned}$$

2. Redefine the prior,

$$w \sim N\left(0,\ \text{diag}\ \left(\sqrt{2}\sigma_w, \sigma_w, ..., \sigma_w\right)\right)$$

3. Take the derivative,

$$\begin{aligned} &-2X^T(y - Xw) + 2\lambda w = 0 \\ &\Rightarrow X^T y = \left(X^T X + \lambda I\right)^{-1} w \\ &\Rightarrow w = \left(X^T X + \lambda I\right)^{-1} X^T y \\ &\Rightarrow w = \frac{1}{1+\lambda} y \end{aligned}$$

For the other problem,

$$\begin{aligned} w &= \left(X^T X + \lambda\ \text{diag}\ \left(\frac{1}{2}, 1, ..., 1\right)\right)^{-1} X^T y \\ \Rightarrow w &= \ \text{diag}\ \left(\frac{1}{1+\frac{1}{2}\lambda}, \frac{1}{1+\lambda}, ..., \frac{1}{1+\lambda}\right) y \end{aligned}$$

4. Use Bernoulli Normal,

$$\begin{aligned} y|w \sim \ \text{Ber}\ (\mu_i), \mu_i = \frac{1}{1+e^{-w^T x_i}}, w \sim N\left(0, \sigma_w^2 I\right) \\ \hat{w} &= \arg\max_w \log p(y|w) + \log p(w) \\ &= \arg\min_w \log\left(\mu_i^{y_i}(1-\mu_i)^{1-y_i}\right) + \lambda\|w\|^2 \\ &= \arg\min_w y_i \log \mu_i + (1-y_i)\log(1-\mu_i) + \lambda\|w\|^2 \end{aligned}$$

$$= \arg\min_w \log\left(1 + e^{-y_i w^T x_i}\right) + \lambda \|w\|^2$$

5. Use singular value decomposition,

$$X = UDV^T, \tilde{X} = U_{[1:r]}$$

36.7 Q7

1. The MLEs are,

$$p_i = \frac{n_i}{n}$$
$$p_{ij} = \frac{n_{ij}}{n_i}$$

2. Use x_k to predict,

$$j = \arg\max_j p_{x_k j}$$

3. Max over both,

$$j = \arg\max_{ij} p_{x_k i} p_{ij}$$

4. Buy stocks.

36.8 Q8

1. Use the one with minimum loss (error)

$$\hat{f} = \arg\min_{f_m} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i, f_m\left(x_i\right)\right)$$

2. The truly best is,

$$\begin{aligned} f^\star &= \arg\min_{f_m} \mathbb{E}\left[l_m\right] \\ \mathbb{P}\left\{\hat{R}\left(f_m\right) - R\left(f_m\right) > t\right\} &\leqslant \mathbb{E}\left[e^{\dfrac{\lambda}{n}\sum_{i=1}^{n} X_i}\right] e^{-\lambda t} \\ &= \mathbb{E}\left[e^{\lambda X_i}\right] e^{-\lambda t} \\ &= e^{\dfrac{\lambda^2}{8n} - \lambda t} \\ &\leqslant e^{-2nt^2}, \lambda = 4nt \end{aligned}$$

Then, union bound,

$$\mathbb{P}\left\{|\hat{R}(f_m) - R(f_m)| > t\right\} \leqslant 2e^{-2nt^2}$$

Union bound again,

$$\begin{aligned}
\mathbb{P}\left\{\bigcup_{f_m} |\hat{R}(f_m) - R(f_m)| > t\right\} &\leqslant 2Me^{-2nt^2} \\
\mathbb{P}\left\{R\left(\hat{f}\right) - \hat{R}\left(\hat{f}\right) < t\right\} &\leqslant 2Me^{-2nt^2} \\
\mathbb{P}\left\{R\left(\hat{f}\right) - \hat{R}(f^\star) < t\right\} &\leqslant 2Me^{-2nt^2} \\
\mathbb{P}\left\{R\left(\hat{f}\right) - R(f^\star) < 2t\right\} &\leqslant 2Me^{-2nt^2}
\end{aligned}$$

3. Given $\hat{R}\left(\hat{f}\right) = 0.1$,

$$\begin{aligned}
\mathbb{P}\{0.1 - R(f^\star) < 2t\} &\leqslant 2Me^{-2nt^2} \\
\delta &= 2Me^{-2nt^2} \\
n &= \frac{1}{2t^2}\log\frac{2M}{\delta}
\end{aligned}$$

36.9 Q9

1. Given $x|w \sim N\left(w, \sigma_v^2\right)$

$$\begin{aligned}
\hat{w} &= \arg\max_w \frac{1}{\sigma_v^2}\|x - w\|^2 \\
&\arg\max_w \|x - w\|^2 \\
&= x \\
\mathbb{E}[w] &= \mathbb{E}[x] \\
&= w \\
\mathbb{V}[w] &= \sigma_v^2
\end{aligned}$$

Same for MSE.

2. Use Gauss-Markov,

$$\begin{aligned}
\hat{w} &= \mathbb{E}[w|x] \\
&= \mathbb{E}[w] + \frac{Cov[w, x]}{\mathbb{V}[X]}(x - \mathbb{E}[x]) \\
&= \frac{\sigma_\theta^2 x}{\sigma_\theta^2 + \sigma_v^2}
\end{aligned}$$

3. This is Quiz 5

36.10 Q10

1. The empirical risk is,

$$\min_{w} \hat{R}(w) = \min_{w} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2$$

2. Approximate by a partition of M

$$\begin{aligned} |w^T x - \tilde{w}^T x| &\leqslant \|w - \tilde{w}\| \|x\| \\ &\leqslant \frac{\sqrt{d}}{M} \end{aligned}$$

   The size of the set is M^d

3. Same as Q8

1. Set δ

$$\begin{aligned} \delta &= 2M^d e^{-2nt^2} \\ M &= \left(\frac{1}{2}\delta e^{2nt^2}\right)^{\frac{1}{d}} \end{aligned}$$

2. Use a smaller space W_s the partitions with at most k non-zero entries.

1. The size of W_s is

$$|W_s| = \sum_{i=0}^{k} \binom{d}{i} (M-1)^i$$

   Then same as Q8

- **Definitions** :

- Bayes Risk is $R^\star = \inf_f R(f) = \inf_f \mathbb{E}[l(f,X,Y)] \overbrace{=}^{0-1loss} \inf_f \mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}\right] = \mathbb{P}\{f(X)\neq Y\}$.

- Optimal Bayes Classifier is $f^\star(x) = 1$ if $\eta(x) \geqslant \frac{1}{2}$ and 0 otherwise; and equivalently $f^\star(x) = 1$ if $\frac{\eta(x)}{1-\eta(x)} \geqslant 1$ and 0 otherwise; and equivalently $f^\star(x) = 1$ if $\frac{p(x|Y=1)\,p(Y=1)}{p(x|Y=0)\,p(Y=0)} \geqslant 1$ and 0 otherwise.

- Log Likelihood Ratio: $\Lambda(x) = \log\left(\frac{p_1(x)}{p_0(x)}\right)$, where $p_j(x) = p(x|Y=j)$.

- Bayes Cost: $C = \sum_{i,j=0}^{1} c_{i,j}\pi_j \mathbb{P}\{\text{ decide } H_i|H_j\} = \sum_{i,j=0}^{1} c_{i,j}\pi_j \int_{R_i} p_j(x)\,dx$, where $\pi_j = \mathbb{P}\{H_j\}$ and $R_j = \{x:\ \text{decide } H_j\}$.

- MLE Risk: $R_{MLE}(q,p_\theta) = \mathbb{E}[-\log p(x|\theta)]$, Excess Risk: $R_{MLE}(q,p_\theta) - R_{MLE}(q,q) = D(q\|p_\theta) \geqslant 0$.

- Probably Approximately Correct: $\mathbb{P}\left\{R\left(\hat{f}\right) - R(f^\star) \leqslant \varepsilon\right\} \geqslant 1-\delta$ where $\varepsilon = \sqrt{\frac{2c^2\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{n}}$, take $\delta = \frac{1}{\sqrt{n}}, \mathbb{E}\left[R\left(\hat{f}\right)\right] \leqslant R(f^\star) + O\left(\sqrt{\frac{\log|\mathcal{F}| + \log n}{n}}\right)$.

- **Estimators** :

- Empirical means and covariances: $\hat{\mu}_j = \frac{1}{\#\{y_i = j\}} \sum_{i:y_i=j} x_i$ and $\hat{\Sigma} = \frac{1}{n}\left(\sum_j \sum_{i:y_i=j} (x_i - \hat{\mu}_j)(x_i-\hat{\mu}_j)^T\right)$.

- Gaussian GLM: $p\left(y|x^Tw\right) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-1}{2}\left(y - x^Tw\right)^2\right)$, and $\hat{w} = \left(X^TX\right)^{-1}X^Ty$.

- Binomial GLM: $p\left(y|x^Tw\right) = \exp\left(y\log\left(\frac{1}{1+e^{-x^Tw}}\right) + (1-y)\log\left(\frac{1}{1+e^{x^Tw}}\right)\right)$.

- Multinomial GLM: $p\left(y|x^Tw\right) = \frac{\exp\left(x^Tw_l\right)}{\sum_{j=1}^{k}\exp\left(x^Tw_j\right)}$.

- Max a Posterior: $\theta_{MAP} = \max_\theta p(\theta|y) \propto p(y|\theta)\,p(\theta)$ to minimize loss $L\left(\theta,\hat{\theta}\right) = \mathbb{1}_{\{\|\hat{\theta}-\theta\|>\varepsilon\}}$.

- Bayesian minimum MSE estimator: $\hat{\theta} = \mathbb{E}[\theta|y]$ to minimize loss $L\left(\theta,\hat{\theta}\right) = \|\theta-\hat{\theta}\|^2$.

- Bayesian minimum MAE estimator: $\hat{\theta} = \text{ median } [\theta|y]$ to minimize loss $L\left(\theta,\hat{\theta}\right) = \|\theta-\hat{\theta}\|_1$.

- Gaussian Penalty Function: $\min_w \log p\left(y|w^Tx\right) + \lambda\|w\|^2$.

- Laplacian Penalty Function: $\min_w \log p\left(y|w^Tx\right) + \lambda\|w\|_1$.

- Sparsity Penalty Function: $\min_w \log p\left(y|w^Tx\right) + \lambda\|w\|_0$.

- Minimax Optimal Estimator: $\hat{\theta} = \arg\min_{\hat{\theta}}\sup_\theta R\left(\hat{\theta},\theta\right)$.

- Support Vector Machine: $\min_{w} \sum_{i=1}^{n} \left(1 - y_i w^T \phi\left(x_i\right)\right)_+ + \lambda \|w\|^2$, or $\min_{w} \|w\|$ such that $\sum_{i=1}^{n} \left(1 - y_i w^T \phi\left(x_i\right)\right)_+ = 0$.

- Exponential Kernel: $K\left(x, x'\right) = \exp\left(x^T x'\right)$.

- Gaussian Kernel: $K\left(x, x'\right) = \exp\left(-\frac{\|x^T - x'\|^2}{\sigma^2}\right)$.

- Laplacian Kernel: $K\left(x, x'\right) = \exp\left(-\beta \|x^T - x'\|\right)$.

- **Distributions** :

- Multivariate Normal Distribution: $p\left(x\right) = \frac{1}{\sqrt{\left(2\pi\right)^d \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(x - \mu\right)^T \Sigma^{-1} \left(x - \mu\right)\right)$ or $\log p\left(x\right) \propto \log\left|\Sigma\right| - \frac{1}{2}\left(x - \mu\right)^T \Sigma^{-1} \left(x - \mu\right)$.

- Binomial Distribution: $p\left(x\right) = \binom{n}{x} p^x \left(1 - p\right)^{n - x}$.

- Hypergeometric Distribution: $p\left(x\right) = \frac{\binom{b}{x}\binom{N - b}{n - x}}{\binom{N}{n}}$.

- Multinomial Distribution: $p\left(x\right) = \binom{n}{x_1 x_2 \ldots x_k} \prod_{i=1}^{k} p_i^{x_i}$.

- Exponential Distribution: $p\left(x\right) = \lambda e^{-\lambda x}$.

- Gamma Distribution: $p\left(x\right) = \frac{x^{a-1} \exp\left(-\frac{x}{b}\right)}{\Gamma\left(a\right) b^a}$.

- Beta Distribution: $p\left(x\right) = \frac{x^{a-1}\left(1 - x\right)^{b-1}}{Beta\left(a, b\right)}$, where Beta(a, b) $= \frac{\Gamma\left(a\right)\Gamma\left(b\right)}{\Gamma\left(a + b\right)}$.

- Exponential Family: $p\left(y|\theta\right) = b\left(y\right) \exp\left(\theta^T T\left(y\right) - \alpha\left(\theta\right)\right)$, θ is the natural parameter and $T\left(y\right)$ is the sufficient statistic. Canonical form is when $T\left(y\right) = y$, and $\log p\left(y|\theta\right) = \sum_{i=1}^{n} \left(w^T x_i y_i - \alpha\left(w^T x_i\right)\right) + \log b\left(y_i\right)$.

- **Other Statistics, Algebra** :

- Kullback-Leibler Divergence: $D\left(p_1 \| p_0\right) = \mathbb{E}_{1\left[\Lambda\left(X\right)\right]} = \int p_1\left(x\right) \log \frac{p_1\left(x\right)}{p_0\left(x\right)} dx$.

- Mahanalobis Distance: (x - $\mu)^T \Sigma^{-1} \left(x - \mu\right)$.

- Sufficiency: $t\left(X\right)$ is sufficient if $p\left(x|t, \theta\right) = p\left(x|t\right)$.

- Rao-Blackwellization: If f is an estimator and t is a sufficient statistic, then $\mathbb{E}\left[f\left(X\right)|t\left(X\right)\right]$ is the improved Rao-Blackwell estimator (in terms of MSE).

- Characteristic Equation of X is $\det(\lambda I - X) = 0$, where λ are the eigenvalues.

- Binomial Conjugate Prior: Binomial(n,p) + Beta(α,β) = Beta$\left(\alpha+\sum_{i=1}^{n}x_i, \beta+n-\sum_{i=1}^{n}x_i\right)$, and Neg Binomal$(r,p)$ + Beta(α,β) = Beta$\left(\alpha+\sum_{i=1}^{n}x_i, \beta+rn\right)$.

- Possion Conjugate Prior: Poisson$(\lambda) + \Gamma(k,\beta)$ = Neg Binomial$\left(k+\sum_{i=1}^{n}x_i, \beta+n\right)$.

- Normal Conjugate Prior: Normal(μ,σ) + Normal(μ_0,σ_0) = Normal$\left(\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\left(\frac{n}{\sigma^2}\bar{x}+\frac{\mu}{\sigma_0^2}\right), \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)$ and multivariate version is, Normal(μ,Σ) + Normal(μ_0,Σ_0) = Normal$\Big(\left(\Sigma_0^{-1}+n\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0+n\Sigma^{-1}\bar{x}\right), \left(\Sigma_0^{-1}+n\Sigma\right.$

- Uniform Conjugate Prior: Uniform$(0,\theta)$ + Pareto(x_m,k) = Pareto$(\max\{x_1,...,x_n,x_m\}, k+n)$.

- Gamma Conjugate Prior: Gamma (α,β) + Gamma (α_0,β_0) = Gamma $\left(\alpha_0+n\alpha, \beta_0+\sum_{i=1}^{n}x_i\right)$.

- Dual Optimization: for $\min_w \sum_{i=1}^{m} l\left(1-y_i x_i^T w\right)+\lambda\|w\|^2$ is $\min_\alpha \sum_{i=1}^{m} l\left(1-y_i\sum_{j=1}^{m}\alpha_j x_i^T x_j\right)+\lambda\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j x_i^T x_j = \min_\alpha \sum_{i=1}^{m} l\left(1-y_i K\alpha\right)+\lambda\alpha^T K\alpha$.

- Shattering Coefficient: for class $\mathcal{F}$ is $\mathcal{S}(\mathcal{F},n) = \max_{x\in X}|\{(f(x_1),...,f(x_n))\in\{-1,+1\}^n, f\in\mathcal{F}\}|$.

- VC Dimension: largest integer k such that $\mathcal{S}(\mathcal{F},k) = 2^k$.

- VC Dimension of Linear Hyperplane: $d+1$.

- VC Dimension of Hyper-Rectangle: $2d$.

- VC Dimension of Neural Net: with sign activation: $O(|E|\log|E|)$, with sigmoid activation: $O\left(|E|^2\right)$.

- **<u>Inequalities, Bounds</u>** :

- Cauchy-Schwarz Inequality: $|\mathbb{E}[XY]| \leqslant \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$ or $|u^T v| \leqslant \|u\|\|v\|$.

- Holder's Inequality: For $\frac{1}{p}+\frac{1}{q}=1, \mathbb{E}[|XY|] \leqslant (E[|X^p|])^{\frac{1}{p}}(E[|X^q|])^{\frac{1}{q}}$.

- Markov's Inequality: For $X \geqslant 0$ and $a > 0, \mathbb{P}\{X > a\} \leqslant \frac{\mathbb{E}[X]}{a}$.

- Chebyshev's Inequality: For $t > 0, \mathbb{P}\{|X-\mathbb{E}[X]| \geqslant t\} \leqslant \frac{\mathbb{V}[X]}{t^2}$.

- Chebyshev-Cantelli Inequality: For $t \geqslant 0, \mathbb{P}\{X-\mathbb{E}[X] > t\} \leqslant \frac{\mathbb{V}[X]}{\mathbb{V}[X]+t^2}$.

- Jensen's Inequality: if f is convex, $\lambda f(x) + (1-\lambda)f(y) \geqslant f(\lambda x+(1-\lambda)y)$, then $\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X])$.

- Association Inequality: if f and g are increasing, $\mathbb{E}\left[f(X)\, g(X)\right] \geqslant \mathbb{E}\left[f(X)\right] \mathbb{E}\left[g(X)\right]$, and if f is increasing, g is decreasing, then the following: $\mathbb{E}\left[f(X)\, g(X)\right] \leqslant \mathbb{E}\left[f(X)\right] \mathbb{E}\left[g(X)\right]$.

- Fourth Moment: $\mathbb{E}\left[|X|\right] \leqslant \left(\mathbb{E}\left[X^2\right]\right)^{\dfrac{3}{2}} \left(\mathbb{E}\left[X^4\right]\right)^{-\dfrac{1}{2}}$.

- Chernoff bound $(1-\delta)$ confidence intervals for mean of $x_i \in [0,1]$ in k dimensions: $\pm\sqrt{\dfrac{\log\left(\dfrac{2k}{\delta}\right)}{2n}}$, and for standard deviation: $\sigma = \dfrac{1}{\sqrt{n}}$. The minimum number of data to ensure ε error with δ probability is $n \geqslant \dfrac{1}{2\varepsilon^2} \log\left(\dfrac{2k}{\delta}\right)$.

- Popoviciu's Inequality: If $\mathbb{P}\{m \leqslant z \leqslant M\} = 1$, then $\mathbb{V}[Z] \leqslant \dfrac{1}{2}(M-m)^2$.

- Hoeffding's Inequality: $X_i \in [a_i, b_i], S_n = \sum_{i=1}^{n} X_i$, for each $t > 0, \mathbb{P}\{|S_n - E[S_n]| \geqslant t\} \leqslant 2\exp\left(-2t^2 \left(\sum_{i=1}^{n} (b_i - a_i)^2\right)^{-1}\right)$.

- Corollary to Hoeffding's Inequality: $X_i \in [a,b], c = (b-a)^2, \hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, then $\mathbb{P}\{|\hat{\mu} - \mu| \geqslant t\} \leqslant 2\exp\left(-\dfrac{2nt^2}{c}\right)$.

- Lemma to proof Hoeffding's Inequality: $\mathbb{E}[Z] = 0, Z \in [a,b]$, then $\mathbb{E}\left[e^{sZ}\right] \leqslant \exp\left(\dfrac{1}{8}\left(s^2 (b-1)^2\right)\right)$.

- Sub Gaussian Tail Bound: if $\{Z_i\}_{i=1}^{n}$ are independent and $\mathbb{P}\{|Z_i - \mathbb{E}[Z_i]| \geqslant t\} \leqslant a\exp\left(-b\dfrac{t^2}{2}\right)$, then $\mathbb{P}\left\{\dfrac{1}{n}\sum_i Z_i - \mathbb{E}[Z] > \varepsilon\right\} \leqslant e^{-cn\varepsilon^2}$ and $\mathbb{P}\left\{\mathbb{E}[Z] - \dfrac{1}{n}\sum_i Z_i > \varepsilon\right\} \leqslant e^{-cn\varepsilon^2}$ with $c = \dfrac{b}{16a}$.

- Gaussian Tail Bound: $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_t^{\infty} \exp\left(-\dfrac{x^2}{2}\right) dx \leqslant \min\left\{\dfrac{1}{2}\exp\left(-\dfrac{t^2}{2}\right), \dfrac{1}{\sqrt{2\pi t^2}}\exp\left(-\dfrac{t^2}{2}\right)\right\}$

- Exponential Bound: If $\mathbb{P}\{|X_i - \mathbb{E}[X_i]| \geqslant t\} \leqslant a\exp\left(-\dfrac{bt^2}{2}\right)$, then $\mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])}\right] \leqslant \exp\left(\dfrac{4as^2}{b}\right)$

- Empirical Risk Minimization: for loss functions bounded by $c, \mathbb{P}\left\{\hat{R}(f) - R(f) > t\right\} \leqslant \dfrac{c^2}{2nt^2}$ by Markov, $\leqslant 2\exp\left(-\dfrac{2nt^2}{c^2}\right)$ by Chernoff and Union bound. If finite, $\mathbb{P}\left\{\hat{R}(f) - R(f) > t\right\} \leqslant \exp\left(-2nt^2\right)$. If multiple $f, \mathbb{P}\left\{\bigcup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| > t\right\} \leqslant 2|\mathcal{F}|\exp\left(-\dfrac{2nt^2}{c^2}\right)$.

- Empirical Risk Comparison: with probability $\delta, R\left(\hat{f}\right) \leqslant \hat{R}\left(\hat{f}\right) + t \leqslant \hat{R}(f^\star) + t \leqslant R(f^\star) + 2t$, where $t = \sqrt{\dfrac{c^2 \log\dfrac{2|\mathcal{F}|}{\delta}}{2n}}$.

- VC Dimension: $\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| > \varepsilon\right\} \leqslant 8\mathcal{S}(\mathcal{F}, n)\exp\left(-\dfrac{n\varepsilon^2}{32}\right)$ and $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|\right] \leqslant 2\sqrt{\dfrac{\log\mathcal{S}(\mathcal{F}, n) + \log 2}{n}}$.

- Uniform Derivation Bound: $\mathbb{P}\left\{\sup_{f\in\mathcal{F}} R(f) - \hat{R}_n(f) \geqslant 1-\delta\right\} \leqslant 6\sqrt{\frac{\mathrm{VC}\ (\mathcal{F})\log\left(\frac{n}{\delta}\right)}{n}}$.
- **<u>Linear Algebra</u>** :
- Singular Value Decomposition: $A = U\Sigma V^T$ satisfy $Av_i = \sigma_i u_i, A^T u_i = \sigma_i v_i$, where $U^TU = UU^T = V^TV = VV^T = I$.
- Schur Complements: $\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I \\ D^- \end{bmatrix}$.
- Matrix Inversion Lemma: (A - $BD^{-1}C)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1} CA^{-1}$.
- Sherman-Morrison Formula: $\left(A^{-1}uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$.
- Vector derivatives: $\frac{dc^Tx}{dx} = c, \frac{dx^Tx}{dx} = 2x, \frac{dx^TAx}{dx} = \left(A + A^T\right)x$.
- Diagonal Metrix: $\left(\Sigma^T\Sigma + \lambda I\right)^{-1}\Sigma^T = \Sigma^T\left(\Sigma^T\Sigma + \lambda I\right)^{-1}$.
- Representation Theorem: $\arg\min_w \sum_{i=1}^n l\left(y_i w^T\phi(x_i)\right) - \lambda\|w\|^2 = \sum_{i=1}^n \alpha_i\phi(x_i)$.
- Kernel Requirement: A kernel is valid iff the Gram matrix $K_{ij} = K(x_i, x_j)$ is symmetric and PSD.
- **<u>Statistics Theorems</u>** :
- weak Law of Large Numbers: $\mathbb{E}\left[|X_i|\right] < \infty \Rightarrow \frac{1}{n}\sum_{i=1}^n X_i \rightarrow^p \mathbb{E}\left[X_i\right]$.
- strong Law of Large Numbers: $\mathbb{E}\left[|X_i|\right] < \infty \Rightarrow \frac{1}{n}\sum_{i=1}^n X_i \rightarrow \mathbb{E}\left[X_i\right]$ as .
- Central Limit Theorem: $\mathbb{E}\left[Z_i\right] = 0, \mathbb{V}\left[Z_i\right] = \sigma^2, \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right) \rightarrow^d N\left(0,\sigma^2\right)$.
- Fisher-Neyman Factorization: $t(X)$ is sufficient iff $p(x|\theta) = a(x)\,b(t,\theta)$.
- Rao-Blackwell Theorem: let $t(X)$ be sufficient and define $g(t(X)) = \mathbb{E}\left[f(X)\,|t(X)\right]$, then $\mathbb{E}\left[\left(g(t(X)) - \theta\right)^2\right] \leqslant \mathbb{E}\left[\left(f(X) - \theta\right)^2\right]$, equal iff $f(X) = g(t(X))$.
- Convergence of Log-Likelihood to KL: $\hat{\theta}_n = \arg\max_\theta p(x|\theta) = \arg\min_\theta \sum_{i=1}^n \log\frac{q(x_i)}{p(x_i|\theta)} \rightarrow \arg\min_\theta D\left(q\|p_\theta\right)$.
- Asyptotic Distribution of MLE: Let $\hat{\theta}_n = \arg\max_\theta p(x|\theta)$, and $\mathbb{E}\left[\frac{\partial\log p(x|\theta)}{\partial\theta}\right] = 0$, then $\hat{\theta}_n \overset{asymp}{\sim} N\left(\theta, n^{-1}I^{-1}(\theta^\star)\right)$ where information matrix $\left[I(\theta^\star)\right]_{j,k} = -\mathbb{E}\left[\left.\frac{\partial\log p(x|\theta)}{\partial\theta_j\partial\theta_k}\right|_{\theta=\theta^\star}\right]$.
- KL-Divergence Information Matrix identity: if $x|\theta \sim N(\theta,\sigma)$, then $\left.\frac{\partial^2 D\left(p(x|\theta)\,\|p(x|\theta^\star)\right)}{\partial\theta^2}\right|_{\theta=\theta^\star} = I(\theta^\star)$.

- Gauss-Markov Theorem: $\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$, then $y|x \sim N\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$, or, $\mathbb{E}[w|x] = \mathbb{E}[w] + \frac{Cov(w,x)}{\mathbb{V}(x)}(x - \mathbb{E}[X])$

- Direct Observation Model: $y = W + \varepsilon, \varepsilon \sim N\left(0, \sigma^2 I\right)$, and the soft-thresholding estimator $\hat{w}_i = \text{sign}(y_i)\max\{|y_i| - \lambda, 0\}$, oracle estimator $\hat{w}_i = y_i \mathbb{1}_{\{|w_i|^2 \geqslant \sigma^2\}}$, with $\lambda = \sqrt{2\sigma^2 \log n}$, then $\mathbb{E}\left[\|\hat{w} - w\|^2\right] \leqslant (2\log n + 1)\left(\sigma^2 + \sum_{i=1}^{n} \min\left\{|w_i|^2, \sigma^2\right\}\right)$.

- **Optimization** :

- Stepsize choice: if $v_t = w_t - w^\star = \left(I - \gamma X^T X\right) v_1$, then $v_t \to 0$ if the eigenvalues of (I - $\gamma X^T X$) $<$ $1 \Rightarrow \gamma < \frac{2}{\lambda\text{max}\left(X^T X\right)}$

- Constant stepsize: If $\|\nabla f_t(w)\| \leqslant G$ and $w^\star = \arg\min_w \sum_{t=1}^{T} f_t(w)$, then gradient descent with $\gamma_t = \gamma$ starting at w_1 satisfies the following: $\frac{1}{T}\sum_{t=1}^{T}\left(f_t(w_t) - f_t(w^\star)\right) \leqslant \frac{\|w_1 - w^\star\|^2}{2\gamma T} + \frac{\gamma}{2}G^2$.

- Diminishing stepsize: If $\|\nabla f_t(w)\| \leqslant G, \|w^\star\| \leqslant B$ and $w^\star = \arg\min_w \sum_{t=1}^{T} f_t(w)$, then gradient descent with $\gamma_t = \frac{1}{\sqrt{t}}$ starting at w_1 satisfies: $\frac{1}{T}\sum_{t=1}^{T}\left(f_t(w_t) - f_t(w^\star)\right) \leqslant \frac{2B^2 + G^2}{\sqrt{T}}$.

- Subgradients: for $\|w\|_1$ is $\text{sign}(w)$; for $\max\left\{0, x^T w\right\}$ is $x\mathbb{1}_{x^T w > 0}$.

- Proimal Gradient: $\min_w f(w) + c(w) \Rightarrow w_k = \text{ prox }\left(w_{k-1} - t\nabla f(w_{k-1})\right), \text{ prox }(v) = \arg\min_u \left(\frac{1}{2}\|u - v\|^2 + tc(u)\right)$.

- 1 Norm Penalty: $\arg\min_w \|y - w\|^2 - \lambda\|w\|_1 = y - \text{sign}(y)\min\{|y|, \lambda\}$.

- 2 Norm Penalty: $\arg\min_w \|y - w\|^2 - \lambda\|w\|^2 = \left(X^T X + \lambda I\right)^{-1} X^T y$.

- General loss funciton: $w = w - 2\mu l'\left(y_i w^T x_i\right) y_i x_i$.

- Perception: $w = w + 2\mu\mathbb{1}_{\{y_i w^T x_i < 0\}} y_i x_i$.

- Backprop: given $y_i^{(l)} = f\left(z_i^{(l)}\right), \frac{\partial J}{\partial y_i^{(l)}} = \frac{\partial J}{\partial y_i^{(l)}} f'\left(z_i^{(l)}\right) y_j^{(l-1)}$, and $\frac{\partial J}{\partial y_i^{(l)}} = \sum_j \frac{\partial J}{\partial y_j^{(l+1)}} f'\left(z_j^{(l+1)}\right) w_{ji}^{(l+1)}, \frac{\partial J}{\partial y_i^{(L)}} = -\left(y_i - y_i^{(L)}\right)$.

- **Other results, formulas** :

- Empirical Classifier Error: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{y}_i \neq y_i\}}$ with mean $\mathbb{E}[\hat{p}] = p$ and $\mathbb{V}[\hat{p}] = \frac{p(1-p)}{n}$, where $p = \mathbb{P}\{\hat{y} \neq y\} = \mathbb{E}\left[\mathbb{1}_{\{\hat{y} \neq y\}}\right]$ is the actual classifier error.

- KL-Divergence of Normal Distribution: With same varaince: $X|Y \sim N(\mu_j, \Sigma)$ with common covariance is $D(p_0\|p_1) = D(p_1\|p_0) = \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0)$. With different variances: $D\left(N(\mu_0, \Sigma_0)\|N(\mu_1, \Sigma_1)\right)$ is $\frac{1}{2}tr\left(\Sigma_1^{-1}\Sigma_0\right) + \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - d + \log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)$.

- Optimal Bayes binary Classifier with common covariances and equal prior is $\hat{y}(x) = 1$ if $2(\mu_1 - \mu_0)^T \Sigma^{-1} x \geqslant \mu_0^T \Sigma \mu_0 - \mu_1^T \Sigma \mu_1$ is linear in x.
- Non-negative Expected Value: If $Y \geqslant 0$, then $\mathbb{E}[Y] = \sum_{i=1}^{\infty} \mathbb{P}\{Y \geqslant i\}$.
- Sum formulas: $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}; \sum_{i=1}^{n} i^2 = \frac{n(n+1)(n+2)}{6}$.
- Bayesian Linear Regression with prior $w \sim N\left(0, \sigma_w^2 I\right)$ has $\hat{w} = \left(X^T X + \lambda I\right)^{-1} X^T y$, where $\lambda = \frac{\sigma^2}{\sigma_w^2}$.
- Minimax Optimal Estimator: if $\hat{\theta}_p = \arg\min_{\hat{\theta}} \int R\left(\hat{\theta}, \theta\right) p(\theta)\, d\theta$, and $\int R\left(\hat{\theta}_p, \theta\right) p(\theta)\, d\theta = \sup_{\theta} R\left(\hat{\theta}_p, \theta\right)$, then $\hat{\theta}_p$ is minimax optimal. In particular, if $R\left(\hat{\theta}_p, \theta\right)$ is constant, then it is minimax.
- Two-layer Neural Net: $\hat{y} = W_2 f(W_1 + b_1) + b_2, W_2 = \alpha, K(x_i, x) = f\left(x_i^T + b_i\right), W_1 = x$ is a SVM.
- Random Feature: if $u_1, ..., u_n, n \geqslant D$ have continuous density, then polynomial mapping $\Phi_n^T = [\phi(u_1), ..., \phi(u_n)]$ has full rank D with probability 1.
- Sauer's Lemma: $\mathcal{S}(\mathcal{F}, n) \leqslant (n+1)^{VC(\mathcal{F})}$