# OLS:

$$h_\theta(x) = \theta^T x$$

$x \in \mathbb{R}^{d+1}$
$\theta \in \mathbb{R}^{d+1}$

ERM with Linear model & Sq. Error Loss

$$X = \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix}_{n \times (d+1)}$$

$$\theta^* = \arg\min_\theta \frac{1}{n} \| X\theta - Y \|_2^2$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

$$(X\theta - Y)^T (X\theta - Y)$$

$$= (X\theta)^T (X\theta) - (X\theta)^T Y - Y^T (X\theta) + Y^T Y$$

$$= Y^T Y + (X\theta)^T X\theta - 2\,\theta^T X^T Y$$

$$\nabla_\theta\, Y^T Y = 0$$

$$\nabla_\theta\, 2\theta^T X^T Y = 2X^T Y$$

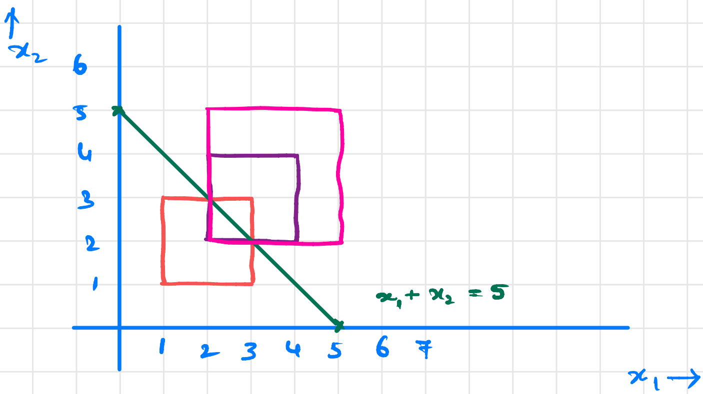
$$\nabla_\theta \left( \theta^T X^T X \theta \right) = 2 X^T X \theta$$

diff wrt $\theta$ & equating to zero.

$$-2X^T Y + 2X^T X \theta \Rightarrow X^T X \theta = X^T Y$$

$$\theta^* = (X^T X)^{-1} X^T Y \blacksquare$$

2. Consider a 2-class PR problem with feature vectors in $\Re^2$. The class conditional density for class-I is uniform over $[1,3] \times [1,3]$ and that for class-II is uniform over $[2,4] \times [2,4]$. Suppose the prior probabilities are equal and we are using 0–1 loss. Consider line given by $x + y = 5$ in $\Re^2$. Is this a Bayes Classifier for this problem? Is Bayes Classifier unique for this problem? If not, can you specify two different Bayes classifiers? Suppose the class conditional densities are changed so that the density for class-I is still uniform over $[1,3] \times [1,3]$ but that for class-II is uniform over $[2,5] \times [2,5]$. Is the line $x + y = 5$ a Bayes classifier now? If not, specify a Bayes classifier now. Is the Bayes classifier unique now? For this case of class conditional densities, suppose that wrongly classifying a pattern into class-I is 10 times more expensive than wrongly classifying a pattern into class-II. Now, what would be a Bayes classifier?




$p_{y=1} = p_{y=2} = 1/2$, 0-1 Loss

Bayes Classifier with 0-1 Loss, is comparing $p_{y=i|x}$

$p_{y=2|x} > p_{y=1|x}$ then $Y = 2$

o/w $Y = 1$

$$p_{Y|X} = \frac{p_{X|Y} \; p_Y}{p_X}$$

Since priors are equal

$p_{X|Y=2} > p_{X|Y=1}$ then $Y=2$

$p_{X|Y=1} \sim U([1,3] \times [1,3])$ $S_1$

Area $(S_1) = 4$

$$p_{X|Y=1}(x) = \begin{cases} \frac{1}{4} & x \in S_1 \\ 0 & \text{otherwise} \end{cases}$$

$p_{X|Y=2} \sim U([2,4] \times [2,4])$ $S_2$

Area $(S_2) = 4$

$$p_{X|Y=2}(x) = \begin{cases} \frac{1}{4} & x \in S_2 \\ 0 & \text{otherwise} \end{cases}$$

$S_1 \cap S_2 = S_{12}$ $[2,3] \times [2,3]$

Area $(S_3) = 1$

$p_{X|Y=1} = p_{X|Y=2} = \frac{1}{4}$

Region A: $S_1 - S_2$

Region B: $S_2 - S_1$

Region C = $S_1 \cap S_2$

Consider $x_1 + x_2 = 5$

$x_1 + x_2 < 5$ then $Y=1$

$x_1 + x_2 > 5$ $Y=2$

Break ties arbitrary

$$R(h_B) = \int_{\mathbb{R}^2} \min\left(p_{y=1}\, p_{x|y=1},\ p_{y=2}\, p_{x|y=2}\right) dx.$$

$$= \int_A + \int_B + \int_C + \int_{\text{outside}}$$

$$= \int_{[2\times3]\times[2\times3]} \frac{1}{8}\, dx = \frac{1}{8}$$

It is easy to see that the Risk involved when we consider the points on $S_1 \cap S_2$ it will be $\frac{1}{8}$

$\therefore$ $x_1 + x_2 = 5$ is a Bayes classifier.

* is it a unique Bayes classifier. : No

Classifier 1 decide $Y=1$ $\forall x \in S_{12}$

Classifier 2 decide $Y=2$ $\forall x \in S_{12}$

<u>**Part 2**</u> :

$$p_{x|y=1} \sim U([1,3]\times[1,3]) \qquad S_1$$

$$p_{x|y=2} \sim U([2,5]\times[2,5]) \qquad S_2$$

$$p_{x|y=1} = \begin{cases} \frac{1}{4} & x \in S_1 \\ 0 & o/w \end{cases}$$

$$P_{X|Y=2} = \begin{cases} 1/9 & x \in S_2 \\ 0 & o/w \end{cases}$$

$$S_1 \cap S_2 = [2,3] \times [2,3]$$

* In $S_{12}$ $\quad p_{X|Y=1} = \frac{1}{4}$ & $p_{X|Y=2} = \frac{1}{9}$

$\frac{1}{4} > \frac{1}{9} \Rightarrow$ in the overlap region we decide $Y=1$

* Now is $x_1 + x_2 = 5$ is it Bayes Classifier.

No, because $x_1 + x_2 > 5 \cap S_{12}$ we say $Y=1$ as per Bayes Classifier but using the line it is $Y=2$

$\therefore x_1 + x_2 = 5$ is not a Bayes Classifier.

* In this case the Bayes Classifier, is Unique.

In the overlap region, No tie $\therefore$ decision is fixed

$\therefore$ its unique.

## Part 3 : Cost Sensitive Bayes classifier.

wrongly classifying the into $Y=1$ is 10 times more expensive than wrongly classifying $Y=2$

$Y=1$, when Truth is $Y=2$ is $\lambda_{12} = 10$

$Y=2$, when truth is $Y=1$ is $\lambda_{21} = 1$

$$\lambda_{11} = \lambda_{22} = 0$$

Risk involve if we decide $Y=1$

$$10\, p_{Y=2|x}$$

Risk involved if we decide $Y=2$

$$1 * p_{Y=1|x}$$

Choose $Y=1$ if

$$10\, p_{Y=2|x} < p_{Y=1|x}$$

Since priors are eque.

$$10\, p_{x|Y=2} < p_{x|Y=1}$$

$$\frac{p_{x|Y=1}}{p_{x|Y=2}} > 10$$

$p_{x|Y=1} = \frac{1}{4}$ $\qquad 10 * p_{x|Y=2} = 10 \times \frac{1}{9} \approx 1.11$

$1.11 < \frac{1}{4}$ $\times$

$\Rightarrow$ In overlap decide $Y=2$

In $S_1 - S_2$ decide $Y=1$

In $S_2 - S_1$ decide $Y=2$

$S_{12}$ decide $Y=2$ due to hight cost of false $Y_1$

6. Suppose $X$ is uniformly distributed over $[0,\ \theta]$, with $\theta > 0$ being the unknown parameter. (The uniform density is given by $f(x) = 1/\theta$, if $0 \le x \le \theta$ and $f(x) = 0$ otherwise). Suppose we have three *iid* samples, $1.75, 0.5, 2.2$. What is the value of the likelihood function $L(\theta|\mathcal{D})$ for (i). $\theta = 10$, (ii). $\theta = 1.9$? Now consider the general case where we represent the three *iid* samples as $x_1, x_2, x_3$. Plot the likelihood function (that is, plot $L(\theta|\mathcal{D})$ versus $\theta$). Now, consider the case where we have $n$ *iid* samples, what is the ML estimate for $\theta$.

$X \sim U[0,\theta]$ with $\theta > 0$, unknown.

$$p_X(x) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & o/w \end{cases}$$

$D = \{ x_1 = 1.75,\ x_2 = 0.5,\ x_3 = 0.22 \}$

Value of likelihood func.

$$L_\theta = p_X(D)$$

$$= \prod_{i=1}^{3} p_X(X = x_i)$$

$$= \prod_{i=1}^{3} \left(\frac{1}{\theta}\right)^3 \ I\,(0 \le x_i \le \theta)$$

The product of Indicator is 1, only when all samples lies in $[0,\theta]$

$0 \le x_1 \le \theta, \quad 0 \le x_2 \le \theta, \quad 0 \le x_3 \le \theta$

$\theta \ge \max\{x_1, x_2, x_3\}$

$$\prod_{i=1}^{3} I\,(0 \le x_i \le \theta) = \begin{cases} 1 & \theta \ge 2.2 \\ 0 & o/w \end{cases}$$

$$L_\theta = \begin{cases} \left(\frac{1}{\theta}\right)^3 & \theta \geq 2.2 \\ 0 & o/w \end{cases}$$

$$L_{\theta=10} = \frac{1}{10^3}$$

$$L_{\theta=1.9} = 0$$

Now in the general case of $n$ samples.

$$L_\theta = \prod_{i=1}^{n} \left(\frac{1}{\theta}\right)^n * I(0 \leq x_i \leq \theta)$$

$$= \left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} I(0 \leq x_i \leq \theta)$$

$$m = \max\{x_1, x_2, \dots, x_n\}$$

$$\prod_{i=1}^{n} I(0 \leq x_i \leq \theta) = \begin{cases} \frac{1}{\theta^n} & \theta \geq m \ \& \ x_i \geq 0 \\ 0 & o/w \end{cases}$$

$$L_\theta = \begin{cases} \theta^{-n} & \theta \geq m \ \& \ x_i \geq 0 \ \forall i \\ 0 & o/w \end{cases}$$

now MLE

maximizer $\theta$, should satisfy $\theta \geq m$

$\theta \in [m, \infty)$

$$L_\theta = \theta^{-n}$$

$\theta \uparrow$, $\theta^{-n} \downarrow$ $\forall n > 0$

so, the max occurs, Only at the smallest allowed point

$\theta = m$

$l(\theta) = \log L_\theta$

$= -n \log \theta$

$\nabla_\theta l(\theta) = \frac{-n}{\theta} < 0 \quad \forall \theta > 0$

so max value is at the left boundary $\theta = m$

$\theta^*_{MLE} = \max\{x_1, x_2, \dots, x_n\}$ ∎

& for our given data

$\theta_{MLE} = 2.2$

4. We want to estimate $\theta$, which is the probability of heads of a coin. The data consists of $N$ tosses of which $N_1$ are heads. Suppose we want a Bayesian estimate. Suppose our prior density is

$$
\begin{aligned}
f(\theta) &= 0.5 \quad \text{if } \theta = 0.5 \\
&= 0.5 \quad \text{if } \theta = 0.6 \\
&= 0 \quad \text{otherwise}
\end{aligned}
$$

**Guess** the MAP estimate for $\theta$ and provide a justification. Then derive the MAP estimate of $\theta$ to verify your intuition.

Estimate $\theta$, prob of heads of a coin.

$N$ - tosses $\quad N_H$ heads $\quad N_T = N - N_H$

$$p(D) = \binom{N}{N_H} \theta^{N_H} (1-\theta)^{N_T}$$

prior is discrete

$$p_\theta = \begin{cases} 0.5 & , \theta = 0.5 \\ 0.5 & , \theta = 0.6 \\ 0 & o/w \end{cases}$$

$$\theta_{MLE} = \frac{N_H}{N}$$

$$p_{\theta|D} \propto p_{D|\theta} \, p_\theta$$

Guess $\theta_{MAP}$:

$$\begin{cases} \theta_{MAP} = 0.6 & \frac{N_H}{N} > 0.5 \\ \theta_{MAP} = 0.5 & o/w. \end{cases}$$

Since $p_{\theta=0.5} = p_{\theta=0.6} = 0.5$

The Comparision $p_{\theta|D}$ is Same as Comparing $p_{D|\theta}$

$$\frac{p_{\theta=0.6|D}}{p_{\theta=0.5|D}} = \frac{p_{D|\theta=0.6} * p_{\theta=0.6}}{p_{D|\theta=0.5} * p_{\theta=0.5}}$$

$$= \frac{p_{D|\theta=0.6}}{p_{D|\theta=0.5}}$$

Choose

$\theta = 0.6$ if $p_{D|\theta=0.6} > p_{D|\theta=0.5}$

Choose

$\theta = 0.5$ if $p_{D|\theta=0.5} > p_{D|\theta=0.6}$

$$p_{D|\theta=0.6} = \binom{N}{N_H} (0.6)^{N_H} (0.4)^{N_T}$$

$$p_{D|\theta=0.5} = \binom{N}{N_H} (0.5)^{N}$$

$$\frac{p_{D|\theta=0.6}}{p_{D|\theta=0.5}} = \frac{(0.6)^{N_H} \quad (0.4)^{N_T}}{(0.5)^{N}}$$

$$= \left(\frac{0.6}{0.5}\right)^{N_H} \left(\frac{0.4}{0.5}\right)^{N_T}$$

$$= (1.2)^{N_H} \quad (0.8)^{N_T}$$

$$p_{D|\theta=0.6} > p_{D|\theta=0.5}$$

$$(1.2)^{N_H} \quad (0.8)^{N-N_H} > 1$$

$$N_H \ln 1.2 + (N - N_H) \ln 0.8 > 0$$

$$N_H \ln 1.2 - N_H \ln 0.8 + N \ln 0.8 > 0$$

$$N_H \ln 1.5 + N \ln 0.8 > 0$$

$$N_H > -N \frac{\ln 0.8}{\ln 1.5}$$

$$N_H > N \frac{\ln (1.25)}{\ln (1.5)}$$

$$N_H > 0.55 N$$

$N_H \gtrsim 0.55 N$ $\qquad \theta_{MAP} = 0.6$

o/w $\qquad \theta_{MAP} = 0.5$ $\qquad \blacksquare$