Mid sem Solution, 2022

Mixed Bag: (A)

For $X_1 > 10$ we get two pure nodes in a single split while for all other options we need more than one splits, thus it will lead to maximum information gain.

Mixed Bag: (B)

True (Discussed in class)

Mixed Bag: (C)

False (Logistic regression and linear regression solve different underlying problem)

Midsem-2022 solution

September 30, 2022

D:

Compared to the variance of the MLE estimate θ_{MLE}, one expects the variance of the MAP estimate θ_{MAP} to be **lower**.
Justification: MAP explicitly takes probability distribution of the data into consideration unlike MLE which maximizes the likelihood only on the basis of training samples. Therefore, it is less likely to overfit the data in comparison to MLE.

E:

For D1, MLE estimate is $\theta_1 = \frac{\sum_{j=1}^{N_1}}{N_1}$
For D2, MLE estimate is $\theta_2 = \frac{\sum_{j=1}^{N_2}}{N_2}$
For D, MLE estimate is $\theta = \frac{\sum_{j=1}^{N_1} + \sum_{j=1}^{N_2}}{N_1+N_2} = \frac{N_1\theta_1 + N_2\theta_2}{N_1+N_2}$
MLE estimate of the coin using D is the weighted average of MLE of D1 and D2.

F:

True
An attribute with binary value will be split at most once in a branch of a tree (based on information gain or gini index).

Midsem Solution Sketch

September 2022

Problem 1

Part G

$$X = \begin{bmatrix} 1 & e^{x_1} & \tan(x_1) \\ 1 & e^{x_2} & \tan(x_2) \\ ... & ... & ... \\ 1 & e^{x_{1000}} & \tan(x_{1000}) \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_{1000} \end{bmatrix}$$

Part H

True, Because no change in the gradient on duplicating the data which implies no change in the weights.
Without duplicating data, the loss function is:-

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2$$

After duplicating data, the loss function is:-

$$J'(\theta) = \frac{1}{2(2n)} \sum_{i=1}^{n} ((h_\theta(x_i) - y_i)^2 + (h_\theta(x_{n+i}) - y_{n+i})^2)$$

$$J'(\theta) = \frac{1}{2(2n)} \sum_{i=1}^{n} ((h_\theta(x_i) - y_i)^2 + (h_\theta(x_i) - y_i)^2)$$

$$J'(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2$$

Part I

$$
\begin{aligned}
P\left(\frac{y=0}{a=(1,0,0,0,0,0)}\right) &= \frac{P\left(\frac{a=(1,0,0,0,0,0)}{y=0}\right).P(y=0)}{P(a=(1,0,0,0,0,0))} \\
&= \frac{P\left(\frac{x_1=1}{y=0}\right).P\left(\frac{x_2=0}{y=0}\right).P\left(\frac{x_3=0}{y=0}\right).P\left(\frac{x_4=0}{y=0}\right).P\left(\frac{x_5=0}{y=0}\right).P\left(\frac{x_6=0}{y=0}\right).P(y=0)}{P(a=(1,0,0,0,0,0))} \\
&= \frac{1.\frac{1}{2}.1.1.\frac{1}{2}.0.\frac{1}{2}}{P(a)} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
P\left(\frac{y=1}{a=(1,0,0,0,0,0)}\right) &= \frac{P\left(\frac{a=(1,0,0,0,0,0)}{y=1}\right).P(y=1)}{P(a=(1,0,0,0,0,0))} \\
&= \frac{P\left(\frac{x_1=1}{y=1}\right).P\left(\frac{x_2=0}{y=1}\right).P\left(\frac{x_3=0}{y=1}\right).P\left(\frac{x_4=0}{y=1}\right).P\left(\frac{x_5=0}{y=1}\right).P\left(\frac{x_6=0}{y=1}\right).P(y=1)}{P(a=(1,0,0,0,0,0))} \\
&= \frac{0.1.\frac{1}{2}.0.1.1.\frac{1}{2}}{P(a)} \\
&= 0
\end{aligned}
$$

Both classes have 0 likelihood probability on $a = (1,0,0,0,0,0)$. So, $a = (1,0,0,0,0,0)$ can be classified randomly either 0 or 1.

Mid Sem Q2 (2022): Solution Sketch

Linear Classifier

Simple Program is as follows:

- Let the input be in the form of a set I. That is

$$I := \{(\mathbf{x_i}, y_i) \mid 1 \leq i \leq N\}$$

- Multiply the x vector of all the points in I with label -1 and collect all the points to pass into `FindPlane` function.

$$P := \{x_i \mid (x_i, y_i) \in I,\ y_i = +1\} \cup \{-x_i \mid (x_i, y_i) \in I,\ y_i = -1\}$$

- Call `FindPlane` on the set of points P. Let w be the hyperplane returned by `FindPlane`.
- If all points in P lie on one side of the hyperplane then return w, otherwise return `Cannot be separated`

Correctness:
If `FindPlane` returns a w such that all points lie on the same side of the plane, then we know that for all points with $y_i = +1$, $w^T x_i > 0$ and for all points with $y_i = -1$, $w^T(-x_i) > 0$. So, for all points with $y_i = +1$ we have $w^T x_i > 0$ and for all points with $y_i = -1$ we have $w^T x_i < 0$. So w is the required vector.
If there is a vector w_0 such that for all points with $y_i = +1$, we have $w_0^T x_i > 0$, and for all points with $y_i = -1$, we have $w_0^T x_i < 0$, then w_0 represents the hyperplane which passes through the origin and all points in the set P (in the program) lie on one side of the hyperplane. So `FindPlane` will return a plane which separates points in I.

Q3

Let there be K mistakes between updates i and j where $w_i = w_j$ $i \neq j$.

Thus, $w_{i+1} = w_i + y_i x_i$

$w_{i+2} = w_{i+1} + y_{i+1} x_{i+1} = w_i + y_i x_i + y_{i+1} x_{i+1}$.

$w_{i+3} = w_{i+2} + y_{i+2} x_{i+2} = w_i + y_i x_i + y_{i+1} x_{i+1} + y_{i+2} x_{i+2}$

. . . .

Thus, $w_j = w_i + \sum_i^k y_t x_t$

∵ $w_i = w_j$,

$w_i = w_i + \sum_i^k y_t x_t$

$\sum_i^k y_t x_t = 0$.

Let X_p = Set of x^i for which $y = +1$
X_N = Set of x_i for which $y = -1$.

$$\sum_i^k x_{pk} - \sum_i^k x_{nk} = 0 \quad \text{--- (1)}$$

If the dataset is linearly separable, $\exists w$ such that $\forall i$ $w^T x_{pi} > 0$ and $w^T x_{ni} < 0$.

Multiply (1) with w^T on both sides,

$$w^T \left(\sum x_{pk} - \sum x_{nk} \right) = 0$$

$$\sum w^T x_{pk} - \sum w^T x_{nk} = 0.$$

$\sum w^T x_{pk}$: positive $\sum w^T x_{nk}$: negative.

Thus LHS > 0 while RHS = 0 ⟶ contradiction.

2020

CS725 Midsem: Solution Sketch

Problem 4: Hold-one-out Regression

The weight vectors $\mathbf{w}, \mathbf{w}_i$ are defined as:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} L(\mathbf{w}) = \arg\min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2 \tag{1}$$

$$\mathbf{w}_i = \arg\min_{\mathbf{w}} L_i(\mathbf{w}) = \arg\min_{\mathbf{w}} ||\mathbf{X}_i\mathbf{w} - \mathbf{y}_i||^2 + \lambda||\mathbf{w}||^2 \tag{2}$$

The difference $L(\mathbf{w}) - L_i(\mathbf{w})$ is:

$$\begin{aligned} L(\mathbf{w}) - L_i(\mathbf{w}) &= ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2 - ||\mathbf{X}_i\mathbf{w} - \mathbf{y}_i||^2 - \lambda||\mathbf{w}||^2 \\ &= ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 - ||\mathbf{X}_i\mathbf{w} - \mathbf{y}_i||^2 \\ &= \Sigma_{\{j\in\mathbb{N};\ j\leq n\}}(\mathbf{x}_j\mathbf{w} - \mathbf{y}_j)^2 - \Sigma_{\{j\in\mathbb{N};\ j\leq n,\ j\neq i\}}(\mathbf{x}_j\mathbf{w} - \mathbf{y}_j)^2 \\ &= (\mathbf{x}_i\mathbf{w} - \mathbf{y}_i)^2 \end{aligned} \tag{3}$$

For $\mathbf{w} = \mathbf{w}_i$,

$$L(\mathbf{w}_i) - L_i(\mathbf{w}_i) = (\mathbf{x}_i\mathbf{w}_i - \mathbf{y}_i)^2 = 0 \tag{4}$$

Now, from equation 2 and 4, we get:

$$\begin{aligned} \mathbf{w}_i &= \arg\min_{\mathbf{w}} L_i(\mathbf{w}) \\ &= \arg\min_{\mathbf{w}} L_i(\mathbf{w}) + (\mathbf{x}_i\mathbf{w} - \mathbf{y}_i)^2 \\ &= \arg\min_{\mathbf{w}} L(\mathbf{w}) \end{aligned} \tag{5}$$

Hence, using the equation 1 and 5, we prove:

$$\mathbf{w}_i = \mathbf{w}^*$$

Q5:

A. Training error remains same. x1 gets a small weight but x2 is sufficient to classify the points (see figure) so training error of 0 should still be achieved.

B. Training error increases. For part B, the decision boundary would be close to a vertical line thus increasing training error.