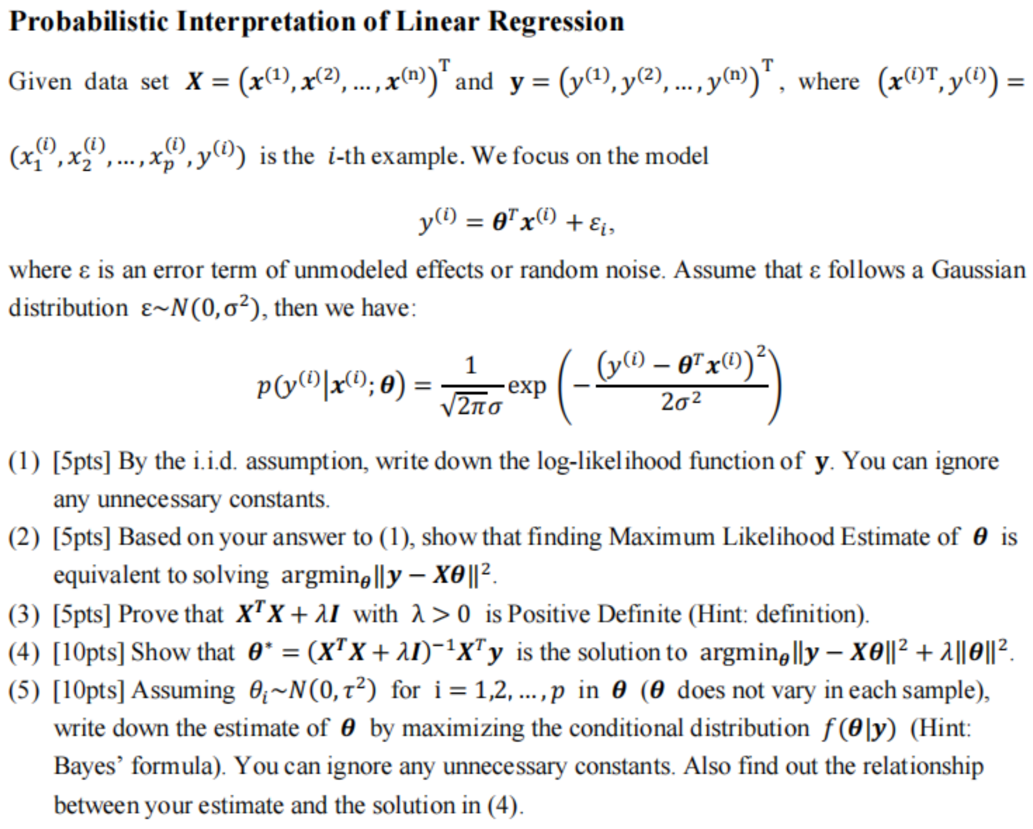
Machine Learning

Assignment 2

Lecture 3 Linear Models

Problem 1



Probabilistic Interpretation of Linear Regression

Given data set $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(\mathrm{n})}\right)^{\mathrm{T}}$ and $\mathbf{y} = \left(y^{(1)}, y^{(2)}, \dots, y^{(\mathrm{n})}\right)^{\mathrm{T}}$, where $\left(\boldsymbol{x}^{(i)\mathrm{T}}, y^{(i)}\right) = (x_1^{(i)}, x_2^{(i)}, \dots, x_p^{(i)}, y^{(i)})$ is the i-th example. We focus on the model

$$y^{(i)} = \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \varepsilon_i,$$

where ε is an error term of unmodeled effects or random noise. Assume that ε follows a Gaussian distribution $\varepsilon \sim N(0, \sigma^2)$, then we have:

$$p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2}{2\sigma^2}\right)$$

(1) [5pts] By the i.i.d. assumption, write down the log-likelihood function of $\mathbf{y}$. You can ignore any unnecessary constants.

(2) [5pts] Based on your answer to (1), show that finding Maximum Likelihood Estimate of $\boldsymbol{\theta}$ is equivalent to solving $\mathrm{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X\theta}\|^2$.

(3) [5pts] Prove that $\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}$ with $\lambda > 0$ is Positive Definite (Hint: definition).

(4) [10pts] Show that $\boldsymbol{\theta}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is the solution to $\mathrm{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$.

(5) [10pts] Assuming $\theta_i \sim N(0, \tau^2)$ for $\mathrm{i} = 1, 2, \dots, p$ in $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ does not vary in each sample), write down the estimate of $\boldsymbol{\theta}$ by maximizing the conditional distribution $f(\boldsymbol{\theta}|\mathbf{y})$ (Hint: Bayes' formula). You can ignore any unnecessary constants. Also find out the relationship between your estimate and the solution in (4).

Solution.

(1)The log-likelihood function of y is:

$$\ell(\boldsymbol{\theta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} lnp(y^{(i)} | \, \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = -nln\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2$$

(2)From the log-likelihood function we can know that:

$$\ell(\boldsymbol{\theta}) \propto -\frac{1}{2\sigma^2} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2$$

Therefore, maximizing $\ell(\boldsymbol{\theta})$ is to minimize $\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2$.

While:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})^2 = \left[y^{(1)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(1)}, \dots, y^{(n)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(n)}\right] \begin{bmatrix} y^{(1)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(1)} \\ \vdots \\ y^{(n)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(n)} \end{bmatrix} = (\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})$$

$$= \|\boldsymbol{y}-\boldsymbol{X\theta}\|^2$$

Therefore, maximum likelihood estimate of $\boldsymbol{\theta}$ is equivalent to solving $argmin_\theta\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2$.

(3)For $\forall \boldsymbol{v}\epsilon R^{p\times 1}$:

$$\boldsymbol{v^T}(\boldsymbol{X^T X} + \lambda \boldsymbol{I})\boldsymbol{v} = \boldsymbol{v^T X^T X v} + \lambda \boldsymbol{v^T I v} = \|\boldsymbol{vX}\|^{\boldsymbol{2}} + \lambda\|\boldsymbol{v}\|^2$$

When $\lambda > 0$, $\|\boldsymbol{vX}\|^{\boldsymbol{2}} + \lambda\|\boldsymbol{v}\|^2 > 0$, i.e. $\boldsymbol{v^T}(\boldsymbol{X^TX}+\lambda\boldsymbol{I})\boldsymbol{v} > \boldsymbol{0}$

Therefore $\boldsymbol{X^TX}+\lambda\boldsymbol{I}$ with $\lambda > 0$ is positive definite.

(4) let

$$\frac{\partial(\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2)}{\partial\boldsymbol{\theta}} = \frac{\partial((\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})+\lambda\boldsymbol{\theta^T\theta})}{\partial\boldsymbol{\theta}} = 2\boldsymbol{X^T}(\boldsymbol{X\theta}-\boldsymbol{y}) + \boldsymbol{2}\lambda\boldsymbol{\theta} = 0$$

We can obtain that $\boldsymbol{\theta}^* = (\boldsymbol{X^TX}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X^Ty}$

While

$$\frac{\partial^2(\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2)}{\partial\boldsymbol{\theta}} = \frac{\partial((\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})+\lambda\boldsymbol{\theta^T\theta})}{\partial\boldsymbol{\theta}} = 2(\boldsymbol{X^TX}+\lambda\boldsymbol{I}) > \boldsymbol{0}$$

i.e. $F(\boldsymbol{\theta}) = \|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2$ is convex

Thus, $\boldsymbol{\theta}^* = (\boldsymbol{X^TX}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X^Ty}$ is the solution to $argmin_\theta\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2$.

(5)From Bayes' formula:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\theta}|\boldsymbol{y})f(\boldsymbol{\theta})}{f(\boldsymbol{y})}$$

Because $f(\boldsymbol{y})$ is a constant, we have:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta},\boldsymbol{y})f(\boldsymbol{\theta})$$

Where

$$f(\boldsymbol{\theta},\boldsymbol{y})f(\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2}{2\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2\tau^2}\right)$$

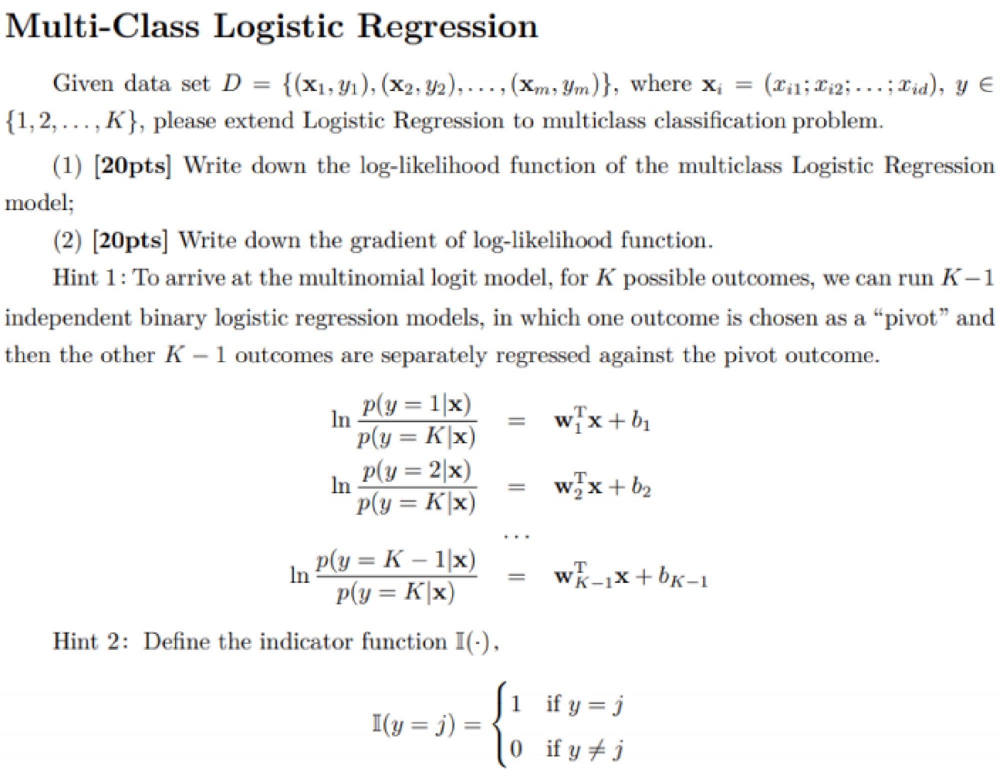
$$\propto \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2}{2\sigma^2}-\frac{\|\boldsymbol{\theta}\|^2}{2\tau^2}\right) = \exp[\frac{1}{2\sigma^2}\left(\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\frac{\sigma^2\|\boldsymbol{\theta}\|^2}{\tau^2}\right)]$$

Thus, maximize $f(\boldsymbol{\theta}|\boldsymbol{y})$ is equivalent to minimize $\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\frac{\sigma^2\|\boldsymbol{\theta}\|^2}{\tau^2}$.

From the solution of (4): the solution to $argmin_\theta\|\boldsymbol{y}-\boldsymbol{X\theta}\|^2+\frac{\sigma^2\|\boldsymbol{\theta}\|^2}{\tau^2}$ is:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X^TX}+\frac{\sigma^2}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X^Ty}$$

Problem 2



Multi-Class Logistic Regression

Given data set $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$, $y \in \{1, 2, \ldots, K\}$, please extend Logistic Regression to multiclass classification problem.

(1) **[20pts]** Write down the log-likelihood function of the multiclass Logistic Regression model;

(2) **[20pts]** Write down the gradient of log-likelihood function.

Hint 1 : To arrive at the multinomial logit model, for K possible outcomes, we can run $K-1$ independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome.

$$\begin{aligned}
\ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^{\mathrm{T}}\mathbf{x} + b_1 \\
\ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^{\mathrm{T}}\mathbf{x} + b_2 \\
&\cdots \\
\ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^{\mathrm{T}}\mathbf{x} + b_{K-1}
\end{aligned}$$

Hint 2: Define the indicator function $\mathbb{I}(\cdot)$,

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{if } y = j \\ 0 & \text{if } y \neq j \end{cases}$$

Solution.

(1)take $p(y = K|\boldsymbol{x})$ as the pivot, we have:

$$p(y = j|\boldsymbol{x}) = \exp\left(\boldsymbol{w}_j^{\boldsymbol{T}}\boldsymbol{x} + b_j\right) \cdot p(y = K|\boldsymbol{x}), with\ j \neq K$$

Normalizing:

$$let \sum_{j=1}^{K-1} p(y = j|\boldsymbol{x}) + p(y = K|\boldsymbol{x}) = p(y = K|\boldsymbol{x})\left(\sum_{j=1}^{K-1} \exp\left(\boldsymbol{w}_j^{\boldsymbol{T}}\boldsymbol{x} + b_j\right) + 1\right) = 1$$

We can obtain that:

$$p(y = K|\boldsymbol{x}) = \frac{1}{\sum_{j=1}^{K-1} \exp\left(\boldsymbol{w}_j^{\boldsymbol{T}}\boldsymbol{x} + b_j\right) + 1}$$

Thus:

$$p(y = j|\boldsymbol{x}) = \exp\left(\boldsymbol{w}_j^{\boldsymbol{T}}\boldsymbol{x} + b_j\right) \cdot p(y = K|\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{w}_j^{\boldsymbol{T}}\boldsymbol{x} + b_j\right)}{\sum_{j=1}^{K-1} \exp\left(\boldsymbol{w}_j^{\boldsymbol{T}}\boldsymbol{x} + b_j\right) + 1}$$

Let $\boldsymbol{\beta} = (\boldsymbol{W}; \boldsymbol{b})$, $\widehat{\boldsymbol{x}} = (\boldsymbol{x}; \boldsymbol{1})$, where $\boldsymbol{W} = (\boldsymbol{w}_1; \boldsymbol{w}_2; \ldots; \boldsymbol{w}_{K-1})$, $\boldsymbol{b} = (b_{\boldsymbol{1}}, b_{\boldsymbol{2}}, \ldots, b_{\boldsymbol{K-1}})$

Then:

$$p(y_i|\boldsymbol{x_i}, \boldsymbol{\beta}) = \sum_{j=1}^{K-1} \mathrm{II}(y_i = j) p(y = j|\widehat{\boldsymbol{x}}_i, \boldsymbol{\beta}) + \mathrm{II}(y_i = K) p(y = K|\widehat{\boldsymbol{x}}_i, \boldsymbol{\beta})$$

$$= \sum_{j=1}^{K-1} \mathrm{II}(y_i = j) \frac{\exp\left(\boldsymbol{\beta_j}^{\boldsymbol{T}}\widehat{\boldsymbol{x}}_i\right)}{\sum_{n=1}^{K-1} \exp\left(\boldsymbol{\beta_n}^{\boldsymbol{T}}\widehat{\boldsymbol{x}}_i\right) + 1} + \mathrm{II}(y_i = K) \frac{1}{\sum_{n=1}^{K-1} \exp\left(\boldsymbol{\beta_n}^{\boldsymbol{T}}\widehat{\boldsymbol{x}}_i\right) + 1}$$

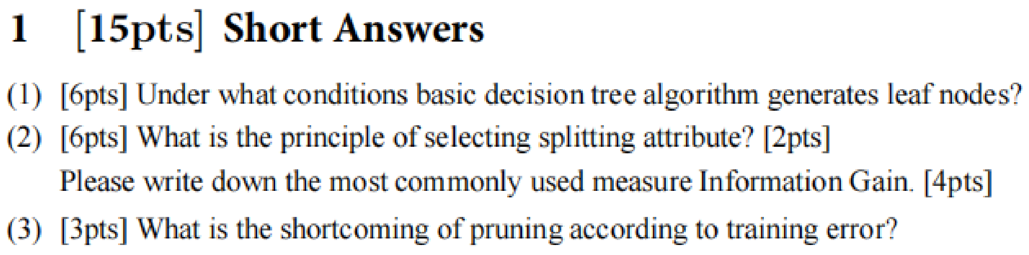
$$\ell(\boldsymbol{\beta})=\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} lnp(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})=\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\left[\sum_{j=1}^{K-1}\mathrm{II}(y_i=j)\,\boldsymbol{\beta_j}^{\boldsymbol{T}}\hat{\boldsymbol{x}}_i-ln\left(1+\sum_{n=1}^{K-1}\exp\left(\boldsymbol{\beta_n}^{\boldsymbol{T}}\hat{\boldsymbol{x}}_i\right)\right)\right]$$

(2)The gradient of log-likelihood function is

$$[\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}]_k=\begin{cases}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\left[\mathrm{II}(y_i=k)\hat{\boldsymbol{x}}_i-\frac{\hat{\boldsymbol{x}}_i\exp\left(\boldsymbol{\beta_n}^{\boldsymbol{T}}\hat{\boldsymbol{x}}_i\right)}{1+\sum_{n=1}^{K-1}\exp\left(\boldsymbol{\beta_n}^{\boldsymbol{T}}\hat{\boldsymbol{x}}_i\right)}\right],k=1,2,\ldots,K-1\\ 0,k=K\end{cases}$$

Lecture 4 Decision Trees

Problem 1



1 [15pts] Short Answers

(1) [6pts] Under what conditions basic decision tree algorithm generates leaf nodes?

(2) [6pts] What is the principle of selecting splitting attribute? [2pts]
Please write down the most commonly used measure Information Gain. [4pts]

(3) [3pts] What is the shortcoming of pruning according to training error?

Solution.

(1) Case1: The attribute set is empty

Case2: All samples from the node belong to the same category

Case3: Attribute values of all the samples from the node are the same

(2) Principle:

Select the attribute with the best classification ability, that is, as the splitting process proceeds, we wish more samples within each node to belong to the same class, i.e. more pure.

Information Gain:

Suppose that the discrete feature a has V possible values $\{a^1,a^2,a^3,\ldots,a^V\}$.

Then, splitting the data set D by feature a produces Vchild node, wherthe vth child node D^v include all samples in D taking the value a^v for feature a.

The information Gain of splitting D with fearture a is:

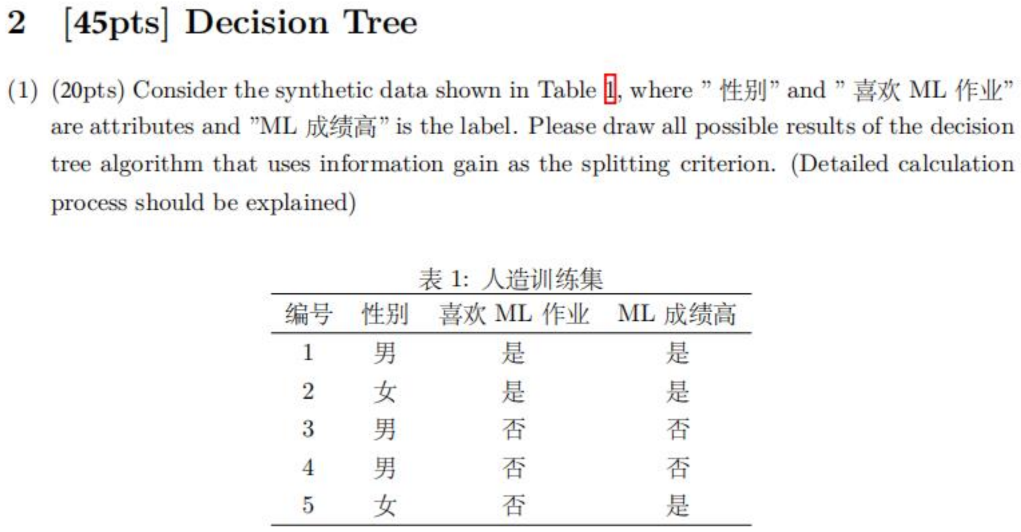
$$Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$$

Where $Ent(D)=-\sum_{i=1}^{c}p_i log_2 p_i$, $|D|$ means the number of samples in data set D.

The higher the information gain, the more purity improvement we can expect by splitting D with feature a.

(3) Cause the risk of overfitting.

Problem 2



2 [45pts] Decision Tree

(1) (20pts) Consider the synthetic data shown in Table 1, where " 性别" and " 喜欢 ML 作业" are attributes and "ML 成绩高" is the label. Please draw all possible results of the decision tree algorithm that uses information gain as the splitting criterion. (Detailed calculation process should be explained)

表 1: 人造训练集

编号	性别	喜欢 ML 作业	ML 成绩高
1	男	是	是
2	女	是	是
3	男	否	否
4	男	否	否
5	女	否	是

Solution.

The attribute set $A = \{$性别，喜欢 ML 作业$\}$

Let the original data set be D_0, after split by 性别 is D_{11}, D_{12}, after split by 喜欢 ML 作业 is D_{21}, D_{22}.

$$Ent(\ D_0) = -0.6log_2 0.6 - 0.4log_2 0.4 = 0.971$$

$$Ent(D_{11}) = 0$$

$$Ent(D_{12)} = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.918$$

$$Ent(D_{21}) = 0$$

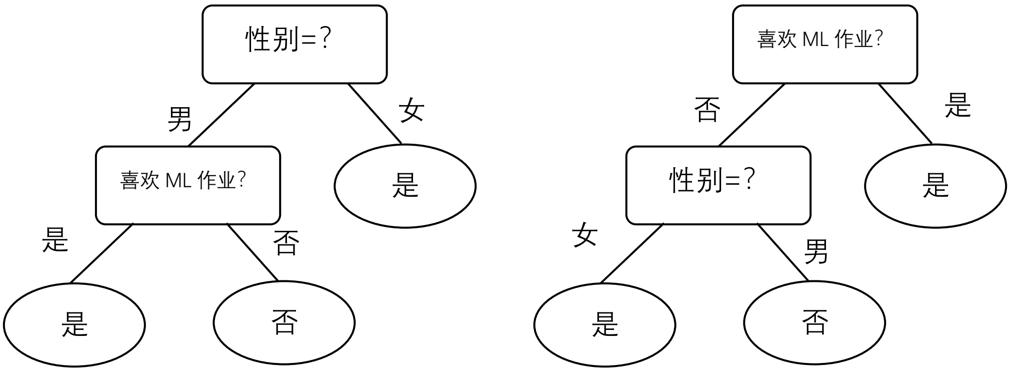
$$Ent(D_{21)} = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.918$$

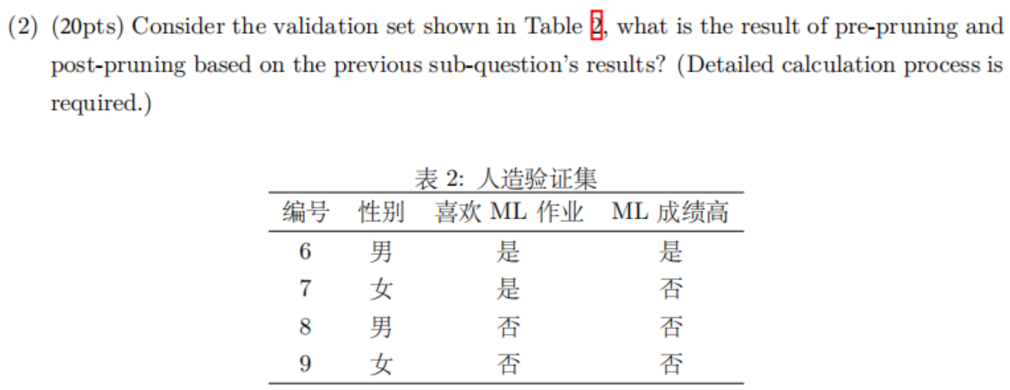
$$Gain(\ D_0, 性别) = Ent(D_0) - \sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v) = 0.42$$

$$Gain(\ D_0, 喜欢ML作业) = Ent(D_0) - \sum_{v=1}^{2}\frac{|D^v|}{|D|}Ent(D^v) = 0.42$$

Therefore, choosing 性别 or 喜欢 ML 作业 firstly are equivalent.

So we have two decision trees as follows:





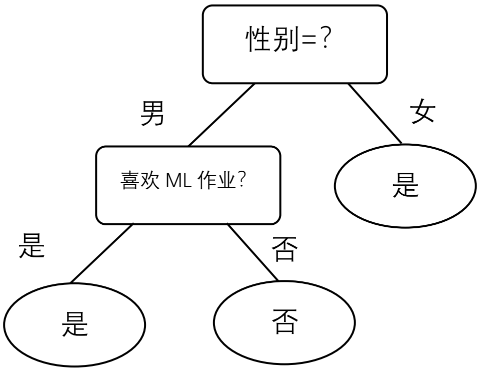
(2) (20pts) Consider the validation set shown in Table 2, what is the result of pre-pruning and post-pruning based on the previous sub-question's results? (Detailed calculation process is required.)

表 2: 人造验证集

编号	性别	喜欢 ML 作业	ML 成绩高
6	男	是	是
7	女	是	否
8	男	否	否
9	女	否	否

【Pre-pruning】

a. For choosing 性别 firstly



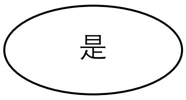
If splitting:

$$Acc_{splitting} = \frac{1}{4} = 25\%$$

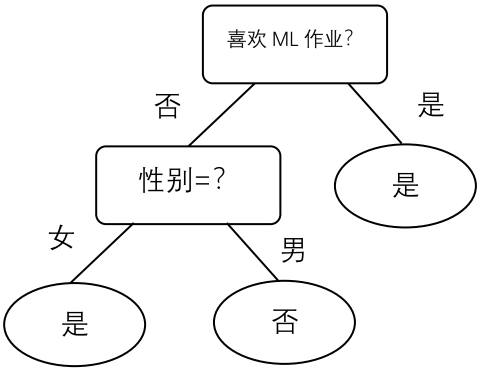
If no splitting:

$$Acc_{no-splitting} = \frac{1}{4} = 25\%$$

So there is no need to split, i.e.:



b. For choosing 喜欢 ML 作业 firstly



If splitting by 喜欢 ML 作业

$$Acc_{splitting} = \frac{3}{4} = 75\%$$

If no splitting by 喜欢 ML 作业

$$Acc_{no-splitting} = \frac{1}{4} = 25\%$$

So it should be split at the first step.

If splitting by 性别

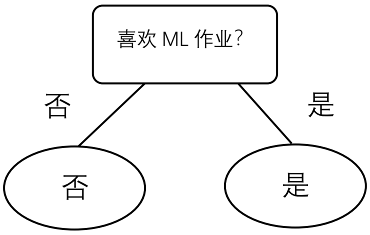
$$Acc_{splitting} = \frac{1}{2} = 50\%$$

If no splitting by 性别

$$Acc_{no-splitting} = \frac{2}{2} = 100\%$$

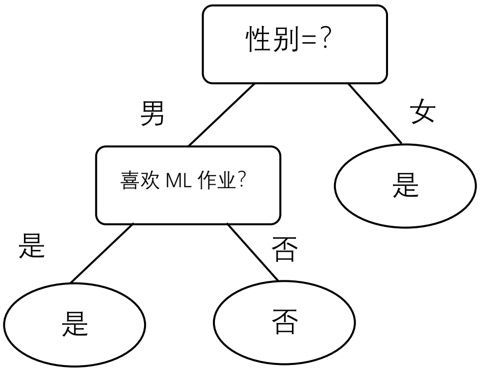
So there is no need to split.

i.e.



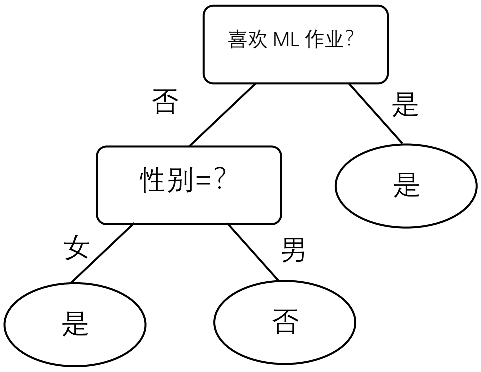
【Post-pruning】

a. For choosing 性别 firstly



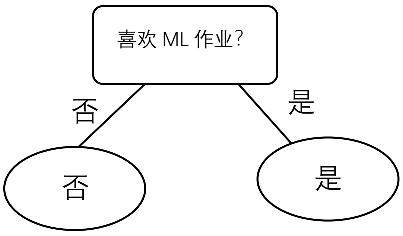
From the bottom up, if we keep 喜欢 ML 作业, the accuracy wil be 50%, otherwise it will be 25%, so it should be kept. So after post-pruning the tree doesn’t change.

b. For choosing 喜欢 ML 作业 firstly



From the bottom up, if we keep 性别, the accuracy wil be 50%, otherwise it will be 75%, so it shouldn't be kept. Then, if we keep 喜欢 ML 作业, the accuracy wil be 75%, otherwise it will be 25%, so it should be kept.

i.e. after post-purning:



(3) (10pts) Compare the results of pre-pruning and post-pruning. What are the accuracies of the two pruning methods on the training set and validation set, respectively? Which method has stronger classification ability?

Solution.

Pruning-method Decision tree	Pre-pruning	Post-pruning
a. version	4/9	7/9
b. version	7/9	7/9

From the comparison, post-pruning is better.