SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Assignment 2

Due Date: March 29, 2024

✏ **Exercise 1**

Log into "cookdata.cn", and enroll the course "数据科学导引". Finish the online exercise there.

✏ **Exercise 2 (Decision Tree)**

You are trying to determine whether a boy finds a particular type of food appealing based on the food's temperature, taste, and size.

Food Sample Id	Appealing	Temperature	Taste	Size
1	No	Hot	Salty	Small
2	No	Cold	Sweet	Large
3	No	Cold	Sweet	Large
4	Yes	Cold	Sour	Small
5	Yes	Hot	Sour	Small
6	No	Hot	Salty	Large
7	Yes	Hot	Sour	Large
8	Yes	Cold	Sweet	Small
9	Yes	Cold	Sweet	Small
10	No	Hot	Salty	Large

1. What is the initial entropy of "Appealing"?
2. Assume that "Taste" is chosen as the root of the decision tree. What is the information gain associated with this attribute.
3. Draw the full decision tree learned from this data (without any pruning).

✏ **Exercise 3: (Maximum Likelihood Estimate (MLE, 极大似然估计))**

Suppose that the samples $\{x_i\}_{i=1}^n$ are drawn from Normal distribution $\mathcal{N}(\mu,\sigma^2)$ with p.d.f. $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$, where $\theta = (\mu,\sigma^2)$. The Maximum likelihood estimator (MLE) of θ is the one that maximize the likelihood function

$$L(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

1. Show that the MLE estimator of the parameters (μ,σ^2) is

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

2. Show that

$$\mathrm{E}\hat{\mu} = \mu, \qquad \mathrm{E}\Big(\frac{n}{n-1}\hat{\sigma}^2\Big) = \sigma^2,$$

   where E is the expectation. This means that $\hat{\mu}$ is an unbiased estimator of μ, but $\hat{\sigma}^2$ is a biased estimator of σ^2.

✏ **Exercise 4 (MLE for Naive Bayes methods)**

Suppose that X and Y are a pair of discrete random variables, i.e., $X \in \{1,2,\ldots,t\}$, $y \in \{1,2,\ldots,c\}$. Then the probability distribution of Y is solely dependent on the set of parameters $\{p_k\}_{k=1}^c$, where $p_k = \Pr(Y=k)$ with $\sum_{k=1}^c p_k = 1$. Similarly, the conditional probability distribution of X given Y is solely dependent on the set of parameters $\{p_{sk}\}_{k=1,\ldots,c}^{s=1,\ldots,t}$, where $p_{sk} = \Pr(X=s|Y=k)$ with $\sum_{s=1}^t p_{sk} = 1$. Now we have a set of samples $\{(x_i,y_i)\}_{i=1}^n$ drawn independently from the joint distribution $\Pr(X,Y)$. Prove that the MLE of the parameter p_k (prior probability) is

$$\hat{p}_k = \frac{\sum_{i=1}^n \mathrm{I}(y_i = k)}{n}, k = 1,\ldots,c;$$

and the MLE of the parameter p_{ks} is

$$\hat{p}_{sk} = \frac{\sum_{i=1}^n \mathrm{I}(x_i = s, y_i = k)}{\sum_{i=1}^n \mathrm{I}(y_i = k)}, s = 1,\ldots,t, k = 1,\ldots,c.$$

✏ **Exercise 5 (Error bound for 1-nearest-neighbor method, optional)** In class, we have estimated that the error for 1-nearest-neighbor rule is roughly twice the Bayes error. Now let us make it more rigorous.

Let us consider the two-class classification problem with $\mathcal{X} = [0,1]^d$ and $\mathcal{Y} = \{0,1\}$. The underlying joint probability distribution on $\mathcal{X}\times\mathcal{Y}$ is $P(\mathbf{X},Y)$ from which we deduce that the marginal distribution of $\mathbf{X}$ is $p_{\mathbf{X}}(\mathbf{x})$ and the conditional probability distribution is $\eta(\mathbf{x}) = P(Y=1|\mathbf{X}=\mathbf{x})$. Assume that $\eta(\mathbf{x})$ is c-Lipschitz continuous: $|\eta(\mathbf{x})-\eta(\mathbf{x}')| \leqslant c\|\mathbf{x}-\mathbf{x}'\|$ for any $\mathbf{x},\mathbf{x}' \in \mathcal{X}$. Recall that the Bayes rule is $f^*(\mathbf{x}) = 1_{\{\eta(\mathbf{x})>1/2\}}$. Given a training set

$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $(\mathbf{x}_i, y_i) \overset{i.i.d.}{\sim} P$ (or equivalently $S \sim P^n$), the 1-nearest-neighbor rule is $f^{1NN}(\mathbf{x}) = y_{\pi_S(\mathbf{x})}$ where $\pi_S(\mathbf{x}) = \arg\min_i \|\mathbf{x} - \mathbf{x}_i\|$.

Define the generalization error for rule f as $\mathscr{E}(f) = \mathrm{E}_{(\mathbf{X},Y)\sim P} 1_{Y \neq f(\mathbf{X})}$. Show that

$$\mathrm{E}_{S\sim P^n} \mathscr{E}(f^{1NN}) \leqslant 2\mathscr{E}(f^*) + c\mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_{\mathbf{X}}} \|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\|.$$

(This means that we can have a precise error estimate for 1-nearest-neighbor rule if we can bound $\mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_{\mathbf{X}}} \|\mathbf{x} - \mathbf{x}_{\pi(\mathbf{x})}\|$.)