Assignment 2

CSE5002 Intelligent Data Analysis (Spring 2022)

Name: Jiang Yuchen ID: 11812419

Q1 MLE and MAP

1. According to the question, $L(\mu) = \prod_{i=1}^{n} p_\theta(x^{(i)}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}) = (\frac{1}{\sigma\sqrt{2\pi}})^n exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2)$. And it doesn't depend on the order of the samples.
2. To get maximum likelihood, set $\frac{d}{d\mu}L(\mu) = (\frac{1}{\sigma\sqrt{2\pi}})^n(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu))exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2)$ as 0. Thus, we can get that $\sum_{i=1}^{n}(x^{(i)}-\mu) = 0 (exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2)$ always greater than zero). As a result, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$.
3. To get maximum a posteriori,
   $\hat{\mu}_{MAP}(x) = \underset{\mu}{argmax} f(\mu|x) = \underset{\mu}{argmax} \frac{f(x|\mu)g(\mu)}{\int_\mu f(x|\mu)g(\mu)d\mu} = \underset{\mu}{argmax} f(x|\mu)g(\mu) = \underset{\mu}{argmax} L(\mu)g(\mu)$
   $= \underset{\mu}{argmax} (\frac{1}{\sigma\sqrt{2\pi}})^n exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2)\frac{1}{\beta\sqrt{2\pi}} exp(-\frac{(\mu-\nu)^2}{2\beta^2}) = \underset{\mu}{argmax}\ exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2 - \frac{(\mu-\nu)^2}{2\beta^2})$
   set $\frac{d}{d\mu}(exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2 - \frac{(\mu-\nu)^2}{2\beta^2})) = (\frac{1}{\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu) + \frac{\nu-\mu}{\beta^2})exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu)^2 - \frac{(\mu-\nu)^2}{2\beta^2}) = 0$, thus, we can get that
   $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x^{(i)}-\mu) + \frac{\nu-\mu}{\beta^2} = 0$. As a result, $\hat{\mu}_{MAP} = \frac{\beta^2\sum_{i=1}^{n} x^{(i)} + \sigma^2\nu}{n\beta^2+\sigma^2}$.
4. As n goes to infinity, $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$ will be the average of all samples $x^{(i)}$ and
   $\hat{\mu}_{MAP} = \frac{\beta^2\sum_{i=1}^{n} x^{(i)} + \sigma^2\nu}{n\beta^2+\sigma^2} = \frac{\beta^2\sum_{i=1}^{n} x^{(i)}}{n\beta^2+\sigma^2} + \frac{\sigma^2\nu}{n\beta^2+\sigma^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x^{(i)}}{1+\frac{\sigma^2}{n\beta^2}} + \frac{\sigma^2\nu}{n\beta^2+\sigma^2}$, when $n \to \infty$, $\frac{\sigma^2\nu}{n\beta^2+\sigma^2} \to 0$ and $1 + \frac{\sigma^2}{n\beta^2} \to 1$.
   Thus, $\hat{\mu}_{MAP} \to \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$, which is same as $\hat{\mu}_{MLE}$.

Q2 Naive Bayes Classifiers

1. According to the question and Bayes' Theorem,
   $p(N|ill) = \frac{p(ill|N)p(N)}{p(ill)} = \frac{2/3*0.5}{0.5} = 2/3$ $p(N|not\ ill) = \frac{p(not\ ill|N)p(N)}{p(not\ ill)} = \frac{1/3*0.5}{0.5} = 1/3$
   $p(C|ill) = \frac{p(ill|C)p(C)}{p(ill)} = \frac{2/3*0.5}{0.5} = 2/3$ $p(C|not\ ill) = \frac{p(not\ ill|C)p(C)}{p(not\ ill)} = \frac{1/3*0.5}{0.5} = 1/3$
   $p(R|ill) = \frac{p(ill|R)p(R)}{p(ill)} = \frac{2/3*0.5}{0.5} = 2/3$ $p(R|not\ ill) = \frac{p(not\ ill|R)p(R)}{p(not\ ill)} = \frac{1/3*0.5}{0.5} = 1/3$
   $p(F|ill) = \frac{p(ill|F)p(F)}{p(ill)} = \frac{1*1/6}{0.5} = 1/3$ $p(F|not\ ill) = \frac{p(not\ ill|F)p(F)}{p(not\ ill)} = \frac{0*1/6}{0.5} = 0$
2. $x^{(2)}$: $P(N, C, \overline{R}, \overline{F}|ill) = 2/3 * 2/3 * (1-2/3) * (1-1/3) = 8/81$,
   $P(N, C, \overline{R}, \overline{F}|not\ ill) = 1/3 * 1/3 * (1-1/3) * (1-0) = 2/27 = 6/81 < 8/81$, $x^{(2)}$ belongs to ill category.
   $x^{(4)}$: $P(N, \overline{C}, \overline{R}, \overline{F}|ill) = 2/3 * (1-2/3) * (1-2/3) * (1-1/3) = 4/81$,
   $P(N, \overline{C}, \overline{R}, \overline{F}|not\ ill) = 1/3 * (1-1/3) * (1-1/3) * (1-0) = 4/27 = 12/81 > 8/81$, $x^{(2)}$ belongs to not ill category.
   $x^{(6)}$: $P(\overline{N}, C, R, \overline{F}|ill) = (1-2/3) * 2/3 * 2/3 * (1-1/3) = 8/81$,
   $P(\overline{N}, C, R, \overline{F}|not\ ill) = (1-1/3) * 1/3 * 1/3 * (1-0) = 2/27 = 6/81 < 8/81$, $x^{(2)}$ belongs to ill category.
   Compared with true labels, $x^{(2)}$ and $x^{(4)}$ are correctly classified, $x^{(6)}$ is not.
3. Similar as 2.,
   $x^{(7)}$: $P(\overline{N}, C, \overline{R}, F|ill) = (1-2/3) * 2/3 * (1-2/3) * 1/3 = 2/81$,
   $P(\overline{N}, C, \overline{R}, F|not\ ill) = (1-1/3) * 1/3 * (1-1/3) * 0 = 0 < 2/81$, $x^{(2)}$ belongs to ill category.
   $x^{(8)}$: $P(N, \overline{C}, \overline{R}, F|ill) = 2/3 * (1-2/3) * (1-2/3) * 1/3 = 2/81$,
   $P(N, \overline{C}, \overline{R}, F|not\ ill) = 1/3 * (1-1/3) * (1-1/3) * 0 = 0 < 8/81$, $x^{(2)}$ belongs to ill category.
   $x^{(9)}$: $P(N, \overline{C}, R, \overline{F}|ill) = 2/3 * (1-2/3) * 2/3 * (1-1/3) = 8/81$,
   $P(N, \overline{C}, R, \overline{F}|not\ ill) = 1/3 * (1-1/3) * 1/3 * (1-0) = 2/27 = 6/81 < 8/81$, $x^{(2)}$ belongs to ill category.

Q3 Ensemble Methods

1. $MSE(\hat{h}) = E[(\hat{h}(x) - h(x))^2] = E[\hat{h}^2(x)] + h^2(x) - 2E[\hat{h}(x)]h(x)$
   $[E_s(\hat{h}(x)) - h(x)]^2 = E_s^2[\hat{h}(x)] + h^2(x) - 2E[\hat{h}(x)]h(x)$
   $E_s[E_s(\hat{h}(x)) - \hat{h}(x)]^2 = E_s[E_s^2(\hat{h}(x)) + \hat{h}^2(x) - 2E_s(\hat{h}(x))\hat{h}(x)] = E_s^2(\hat{h}(x)) + E_s[\hat{h}^2(x)] - 2E_s^2(\hat{h}(x)) = E_s[\hat{h}^2(x)] - E_s^2(\hat{h}(x))$
   Thus,
   $[E_s(\hat{h}(x)) - h(x)]^2 + E_s[E_s(\hat{h}(x)) - \hat{h}(x)]^2 = (E_s^2[\hat{h}(x)] + h^2(x) - 2E[\hat{h}(x)]h(x)) + (E_s[\hat{h}^2(x)] - E_s^2(\hat{h}(x))) = E_s[\hat{h}^2(x)] + h^2(x) - 2E[$
2. Because bagging method generate aggregated results over multiple sets of data and reduce the variance, but Naive Bayes classifiers is not low-bias classifiers with high variances. In Naive Bayes classifiers, we use whole datasets to reflect the character of each variable. However, bootstrap sampling decrease the information of training sets. As a result, it's normally hard to improve the performance.

3. Similarly as 2., Random Forests is low-bias classifiers with high variances, which means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees. Bootstrap sampling is a way of de-correlating the trees by showing them different training sets. Thus, bagging method can improve the performance.

Q4 Model Assessment

1. Confusion matrix is a table which compares predictions with ground truth to show the performance of the model. There are four categories: True Positive, True Negative, False Positive and False Negative.
2. ROC curve is Receiver Operating Characteristic Curve. It's generated by varying the threshold of the classifier and get the performance of the classifier. AUC is Area Under ROC Curve.

   We use ROC curve to measure the performance of binary classifier as threshold is varied and it gives full characteristic of the classifier in terms of sensitivity (TP rate) vs. specificity (1 − FP rate).

   AUC ranges from [0,1] and higher AUC means higher accuracy .AUC will often be a better classifier's evaluation metric than accuracy (thresholding at 0), especially for imbalanced data and unknown misclassification costs.