Machine Learning

Assignment 1

Fatemeh Nadi 810101285

November 10, 2022

Problem 1

Using the books notation where $\lambda_{\kappa i}$ is the loss associated with us deciding an object is from class i when it in fact is from class κ.

$$\frac{p\left(x|w_1\right)}{p\left(x|w_2\right)} \underset{<_1}{\overset{>^2}{}} \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}} \cdot \frac{p\left(\omega_2\right)}{p\left(\omega_1\right)}$$

Assuming that $\lambda_{11}=\lambda_{22}=0$ and x_0 is where this equation becomes an equality so:

$\frac{p(x|w_1)}{p(x|w_2)} = \frac{\lambda_{21}}{\lambda_{12}} \cdot \frac{p(\omega_2)}{p(\omega_1)}$

The general form of probability density function for a normal distribution:

$$\mathbf{N}\left(x|\mu,\sigma^2\right) \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} \tag{1}$$

and we assume:

$p\left(x|w_1\right) \sim \boldsymbol{N}\left(0,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$,

$p\left(x|w_2\right) \sim \boldsymbol{N}\left(1,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-1}{\sigma}\right)^2}$

so:

$\frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-1}{\sigma}\right)^2}} = \frac{\lambda_{21}}{\lambda_{12}} \cdot \frac{p(\omega_2)}{p(\omega_1)} \rightarrow e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} = \frac{\lambda_{21}p(\omega_2)}{\lambda_{12}p(\omega_1)} \cdot e^{-\frac{1}{2}\left(\frac{x-1}{\sigma}\right)^2}$

We take the logarithm from both sides of the equation:

$\ln \frac{\lambda_{21}p(\omega_2)}{\lambda_{12}p(\omega_1)} - \frac{x^2-2x+1}{2\sigma^2} = -\frac{x^2}{2\sigma^2}$,

$\ln \frac{\lambda_{21}p(\omega_2)}{\lambda_{12}p(\omega_1)} = \frac{-2x+1}{2\sigma^2}$

$$\boxed{x = \frac{1}{2} - \sigma^2 \ln \frac{\lambda_{21}p\left(\omega_2\right)}{\lambda_{12}p\left(\omega_1\right)}}$$

Problem 2

A.

Let us recall from the probability course basics the Bayes rule:

$$p\left(x,\omega_i\right)=p\left(x|\omega_i\right)p\left(\omega_i\right) \tag{2}$$

Where $p\left(x\right)$ is the pdf of x and for which we have and M is number of classes:

$$p\left(x\right)=\sum_{i=1}^{M}p\left(x|\omega_i\right)\ p\left(\omega_i\right)$$

$$\begin{aligned} p\left(x\in R_i,\omega_i\right) &\overset{2}{=} p\left(x\in R_i,\omega_i\right)=p\left(x\in R_i|\omega_i\right)\ p\left(\omega_i\right) \\ &=\int_{R_i}p\left(x|\omega_i\right)\ p\left(\omega_i\right) \end{aligned} \tag{3}$$

The probability of correct classification of PC is defined:

$$\begin{aligned} p_{correct} &=\sum_{i=1}^{M}p\left(x\in R_i,\omega_i\right) \\ &\overset{3}{=}\sum_{i=1}^{M}\left(\int_{R_i}\left(p\left(x_i|\omega_i\right)dx\right)p\left(\omega_i\right)\right) \\ &=\sum_{i=1}^{M}\left(\int_{R_i}p\left(x_i|\omega_i\right)p\left(\omega_i\right)dx\right) \end{aligned}$$

Bayesian classifier select ω_i in this situation:

$$p\left(x_i|\omega_i\right)p\left(\omega_i\right)>p\left(x_j|\omega_j\right)p\left(\omega_j\right)\quad\forall i\neq j$$

So $p_{correct}$ is maximized as large as possible and Bayesian classifier is optimal for multi classes.

B.

We know that: $\sum_{i=1}^{M}p\left(\omega_i|x\right)=1$ and $p_e=1-\underset{i}{arg\ max}\ p\left(\omega_i|x\right)$
The minimum probability of a class is $\frac{1}{M}$, and if it is less than this value, the sum of the probabilities does not become 1 so:

$$p\left(\omega_i|x\right)\leq\frac{1}{M}$$

The lowest value that $\underset{i}{arg\ max}\ p\left(\omega_i|x\right)$ can take is $\frac{1}{M}$ so:

$$p_e = 1 - \underset{i}{\arg\max}\, p\left(\omega_i|x\right) \leq 1 - \frac{1}{M}$$

$$\boxed{p_e \leq \frac{M-1}{M}}$$

C.

ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.
Area under the ROC Curve (AUC) provides an aggregate measure of performance across all possible classification thresholds, the Higher the AUC, the better the model is at distinguishing between classes, ROC curve plots TPR vs. FPR at different classification thresholds, but this concept is not immediately applicable for muticlass classifiers.

A solution for this problem is we can take one class and consider it as our "positive" class, while all the others (the rest) are considered as the "negative" class. by doing this, we reduce the multiclass classification output into a binary classification one, and so it is possible to use all the known binary classification metrics to evaluate, we must repeat this for each class present on the data.

For example, for a 3-class data as $\omega_1, \omega_2, \omega_3$, consider ω_1 as the desired class that data is in this class or not, so the main problem became a two-class subset, now that the problem is binary we can also use the same metrics we use for binary classification.

D.

the assumption in Naïve Bayes is that features are conditionally independent given the predicted variable, not independent.
Note also that, even though this simplification makes naïve assumptions about the conditional joint distribution of features that are in many cases far from the true distribution for that purpose, our simplification strategy may be not good enough.
Anyway if our features are conditionally independent therefore Naïve Bayes is optimal, otherwise, we must use an optimal bayesian classifier.
However, independence of features doesn't necessarily imply conditional independence, given only general independence of features.

Problem 3

Should show that:

$$\widehat{\mu}_{MAP} = \frac{z + \sqrt{z^2 + 4R}}{2R}$$

Assuming that:

$$z = \frac{1}{\sigma^2} \sum_{k=1}^{N} x_k \tag{4}$$

$$R = \frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2} \tag{5}$$

According to the assumptions $X \sim N(\mu, \sigma)$ and pdf μ:

$$p(\mu) = \frac{\mu exp\left(\frac{-\mu^2}{2\sigma_\mu^2}\right)}{\sigma_\mu^2} \tag{6}$$

We know likelihood function is just the joint pdf of data so:

$$Likelihood(\mu) = \left(\prod_{i=1}^{N} p\left(x_i|\mu, \sigma^2\right)\right) p(\mu) \tag{7}$$

We work with the log-likelihood because some computations are easier

$$\begin{aligned}
\ln\left(Likelihood(\mu)\right) &= \sum_{i=1}^{N} \ln\left(p\left(x_i|\mu, \sigma^2\right)\right) + \ln\left(p(\mu)\right) \\
&\overset{1}{=} \ln\left(p(\mu)\right) + \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)\right) \\
&\overset{6}{=} \ln\left(\frac{\mu \cdot exp\left(\frac{-\mu^2}{2\sigma_\mu^2}\right)}{\sigma_\mu^2}\right) + \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)\right) \\
&= \ln(\mu) - \frac{\mu^2}{2\sigma_\mu^2} + \ln\left(\sigma_\mu^2\right) + \sum_{i=1}^{N} \left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \left(\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)\right) \\
&= \ln(\mu) - \frac{\mu^2}{2\sigma_\mu^2} + \ln\left(\sigma_\mu^2\right) + N \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{N} (x-\mu)^2
\end{aligned}$$

Now to maximize $Likelihood\,(\mu)$ take the derivative with respect to μ:

$$\frac{\partial Likelihood\,(\mu)}{\partial \mu} = \frac{1}{\mu} - \frac{\mu}{\sigma_\mu^2} + 0 + 0 + \frac{1}{\sigma^2}\sum_{i=1}^{N}(x-\mu) = 0$$

$$\frac{1}{\mu} - \frac{\mu}{\sigma_\mu^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i - \frac{N\mu}{\sigma^2} = 0$$

$$\overset{5}{\rightarrow} \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2}\right)\mu - R = 0$$

$$\overset{4}{\rightarrow} z\mu - R - \frac{1}{\mu} = 0$$

$$\overset{\times\mu}{\rightarrow} z\mu^2 - R\mu - 1 = 0$$

Solve Quadratic Equations using the Quadratic Formula:
Evaluate $\Delta = b^2 - 4ac$, $x = \frac{-b \pm \sqrt{\Delta}}{2a}$ when $a = z, b = -R, c = -1$ and $x = \mu$:

$\Delta = R^2 + 4Z$

$$\boxed{\mu = \frac{z \pm \sqrt{z^2+4R}}{2R} = \widehat{\mu}_{MAP}}$$

Problem 4

A.

According to the mentioned assumptions we have $x_1, x_2, ..., x_d$ iid Bernoulli(ρ)

$X_i \sim Ber(p)$

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)} \tag{8}$$

We want to find out what that p is.
Ask for MLE for p so we are going to use MLE to estimate the p parameter of a Bernoulli distribution , its often easier to work with the log-likelihood in these situations than the likelihood. Note that the maximum of the log-likelihood is exactly the same as the max of the likelihood.

$\ell(p) = \ln p \sum_{i=1}^{n} x_i + \ln(1-p) \sum_{i=1}^{n}(1-x_i)$

let's compute the derivative:

$\frac{\partial \ell(p)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p} \overset{\text{set}}{=} 0$

$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i = p\sum_{i=1}^{n}(1-x_i)$

$$\boxed{\text{p} = \frac{1}{n}\sum_{i=1}^{n} x_i} \tag{9}$$

B.

consistency of MLE
$\widehat{p}$ is a estimator using $\{x_1, x_2, ..., x_d\}$, we say that $\widehat{p_d}$ is consistent if $\widehat{p_d} \xrightarrow{p} p$,
if $\Theta = E\left[\widehat{\theta}\right] \rightarrow$ this estimator is unbiased.
If this function wants to have a probability of behavior equal to its actual value, its expected value should be equal to itself at infinity, and given a big enough sample size, you can expect that your estimator will be close to the true value of the parameter,as a result:

$$E\left[\widehat{p}\right] \overset{9}{=} \int_{-\infty}^{+\infty} \left(\frac{1}{d}\sum_{i=1}^{d} x_i\right) p\left(x_i|\Theta\right) dx$$

We know $p(\omega_1) = p(\omega_1) = \frac{1}{2}$ therefore half of space is equal to $1 \rightarrow \sum_{i=1}^{d} x_i = \frac{d}{2}$

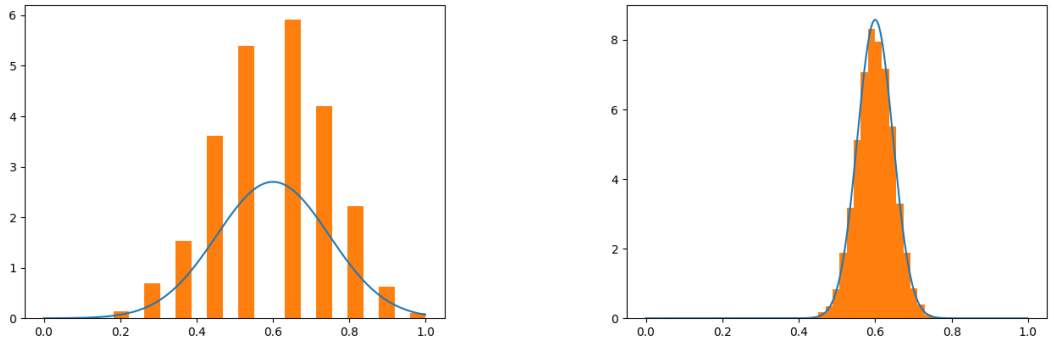
$$\lim_{d\to\infty} E\left[\widehat{p}\right] = \int_{-\infty}^{+\infty} \left(\frac{1}{d}\cdot\frac{d}{2}\right) p\left(x_i|\Theta\right) dx \simeq p$$

Since we are at infinity, we can ignore the factor of 1/2.

C.

Please see this file: ”Q4.ipynb”
Code and explanation are provided.

Asymptotic Normality



(a) d=11 (b) d=111

Figure 1: distribution of $p(T|\omega_i)$

You can look at it this way : when your sample size is big enough, the distribution of the parameter estimator can be approximated by a normal distribution(Asymptotic Normality) with variance which is decreasing with the sample size. The bigger the sample size, the smaller the variance in the limit, the variance vanishes completely, hence, convergence to a constant is achieved.
$p\left(T|\omega_i\right)$ has the asymptotic distribution, $\widehat{p} = T \xrightarrow{d} N\left(p, \sigma^2\right)$

Problem 5:

A.

The Bayes Optimal Classifier is a probabilistic model that makes the most probable prediction for a new example using the training data and space of hypotheses and use the bayes theorem whitch is a method for calculating a hypothesis's probability based on its prior probability, the probabilities of observing specific data given the hypothesis, and the seen data itself.
Naive Bayes classier assumes conditional independence among the attributes. It assumes that $p(x, y) = p(x)\, p(y)$ or in other words, $p(x|y, z) = p(x|z)$, whereas Optimal Bayes classier does not make such an assumption. It is obvious that Optimal Bayes Classier is much more complex and computationally more expensive.

Please see this file: "Q5.ipynb"
Code and explanation are provided.

B.

Accuracy of my Naive Bayes classier are 92.1%.

Class 0 F1 Score: 0.88
Class 1 F1 Score: 0.94

Class 0 Precision: 0.97
Class 1 Precision: 0.90

Class 0 Recall: 0.81
Class 1 Recall: 0.97

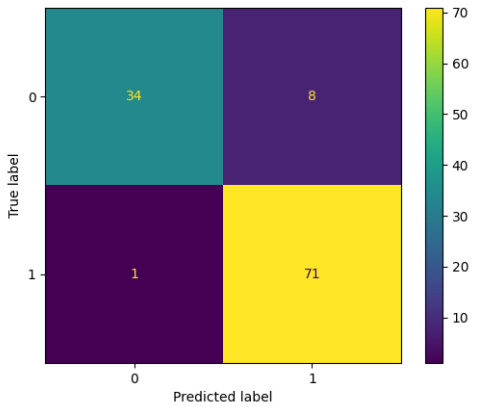


Figure 2: Confusion Matrix for my Naive Bayes classier

C.

Please see this file: ”Q5.ipynb”
Code and explanation are provided.
Accuracy of SKLearn’s Naive Bayes Classier are 91.22%.

Class 0 F1 Score: 0.87
Class 1 F1 Score: 0.93

Class 0 Precision: 0.94
Class 1 Precision: 0.90

Class 0 Recall: 0.81
Class 1 Recall: 0.97

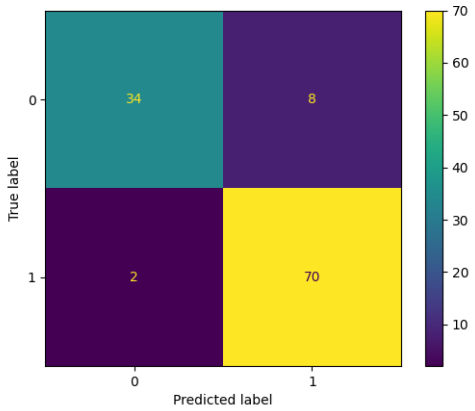


Figure 3: Confusion Matrix for SKLearn’s Naive Bayes Classier

The accuracy of my Naive Bayes classier is 92% and the accuracy of SKLearn’s Naive Bayes Classier is 91%, my model is better than sklearn library and predicts one sample in class 1 more than the other.

Problem 6

Please see this file: ”Q6.ipynb”
Code and explanation are provided.

According to figure 4 model correctly classifies 57 Manchester and 46 Chelsea photos, and when dividing this to the whole number of the photos we get 84% accuracy.
This model classifies Manchester photos more correctly and tends to misclassify Chelsea photos as Manchester.
therefore, Manchester has a lower precision and higher recall than Chelsea.

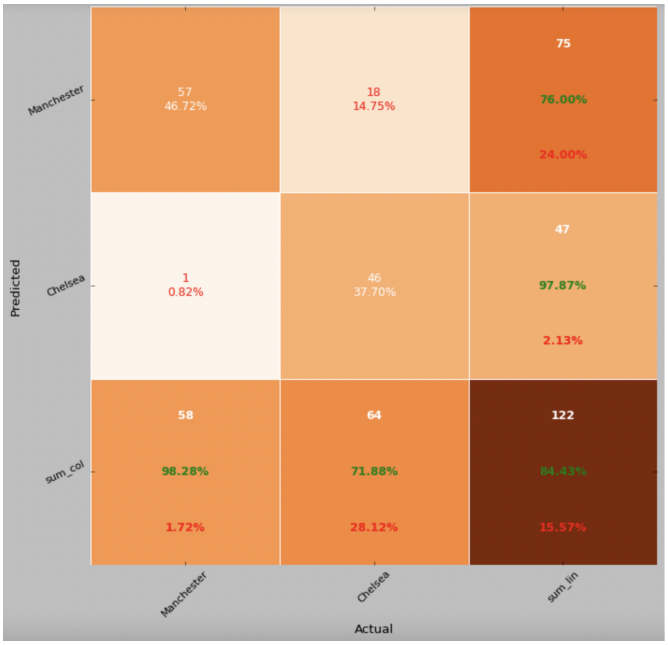


Figure 4: Confusion Matrix for our model

Manchester results:
Precision: 0.97
Recall: 0.71
F1 score: 0.82