Assignment 1

Leroy Souz

19th September 2024

Question 1

1.1 If precision rate is used to measure the empirical error on the training set, what kind of circle classifier can be trained? Please provide an example.
Precision rate counts the accuracy of positives predicted by the model. The formula used is $\frac{TP}{TP+FP}$.
A circle classifier being trained to maximize precision would be one that tries to minimize the amount of negative samples in its area while maximizing positive samples. One of such circle could be one that has only 1 postive sample and no negative samples. This would give a precision rate of 1.

1.2 If recall rate is used to measure the empirical error on the training set, what kind of circle classifier can be trained? Please provide an example.
Recall rate calculates how good the model is at identifying positives predictions. The formula used is $\frac{TP}{TP+FN}$.
A circle classifier being trained to maximize recall would try to maximize the positive examples in its area as possible which ignoring negative samples being included as these wouldn't have any effect on the recall rate. One of such circle could be one that has all positive samples. This would give a recall rate of 1.

1.3 What problems are expected for the case of 1.1 and 1.2? Would you suggest another metric to learn a reasonable circle?
When using either one of precision or recall, the model can become really biased to either favouring positive accuracy or positive samples which can lead to in accurracte predictions. To counter this problem, a mix of both precision and recall can be used to make the model balance between both. This can be done using the F1 rate whose formla is: $\frac{2*precision*recall}{precision+recall}$.

Question 2

2.1 Define the sample spaces $\Omega_2, \Omega_3, \cdots \Omega_m$ for the following experiments: At every level, the ball only has two options paths to choose from, left or right. The sample space for each level is $\Omega_i = \{L, R\}$.

Therefore, the sample space for each level will be:

$$\begin{aligned}
\Omega_2 &= \{L, R\} \\
\Omega_3 &= \{L, R\} \\
&\vdots \\
\Omega_m &= \{L, R\}
\end{aligned}$$

2.2 Define the sample space $\Omega = \Omega_2 \times \Omega_3 \times \cdots \Omega_m$. Each element of the set Ω encodes a path the ball could take until it arrives at the ground-level L_G
The sample space Ω is the cartesian product of all the sample spaces Ω_i for each level. Therefore, the sample space Ω is:

$$\begin{aligned}\Omega &= \Omega_2 \times \Omega_3 \times \cdots \Omega_m \\ &= \{L, R\} \times \{L, R\} \times \cdots \{L, R\} \\ &= \{(L, L, \cdots, L), (L, L, \cdots, R), \cdots, (R, R, \cdots, R)\}\end{aligned}$$

2.3 What is the meaning of the location at L_G where the ball finally arrives? (hint: think about the possible ball path resulting in the different locations like the leftmost (blue star) or the second left location (green star)).
If we label the leftmost spot as 0, the location L_G where the ball finally arrives is the number of right paths taken by the ball.

2.4 How would you represent the location numerically? Please define a random variable that map Ω to the numerical values.
Let X be the random variable that maps Ω to the numerical values. The random variable X is defined as:

$$X(\Omega) = \text{number of right paths taken by the ball}$$

2.5 Define the PMF of your random variable for the depth M = 5, M = 10, M=100. Plot them and check how the PMFs change as M goes to large. Please explain the phenomenon in relation to Central Limit Theorem.
The PMF of the random variable X is:

$$P(X = x) = \binom{M}{x} \cdot p^x \cdot (1-p)^{M-x}$$

When M = 5, the PMF is:

$$\begin{aligned}P(X = x) &= \binom{5}{x} \cdot 0.5^x \cdot 0.5^{5-x} \\ &= \binom{5}{x} \cdot 0.5^5\end{aligned}$$

When M = 10, the PMF is:

$$\begin{aligned}P(X = x) &= \binom{10}{x} \cdot 0.5^x \cdot 0.5^{10-x} \\ &= \binom{10}{x} \cdot 0.5^{10}\end{aligned}$$

When M = 100, the PMF is:

$$\begin{aligned}P(X = x) &= \binom{100}{x} \cdot 0.5^x \cdot 0.5^{100-x} \\ &= \binom{100}{x} \cdot 0.5^{100}\end{aligned}$$

As M goes to large, the PMF becomes more and more like a normal distribution. This is because of the Central Limit Theorem which states that the sum of a large number of independent random variables will be approximately normally distributed.

Question 3

We are given that $P(D|W) = 0.20$, $P(D|W') = 0.80$, and $P(W') = 0.30$

3.1 What's the chance that your plant will survive the week?

$$\begin{aligned} P(D) &= P(D|W)P(W) + P(D|W')P(W') \\ &= 0.20 * 0.70 + 0.80 * 0.30 \\ &= 0.14 + 0.24 \\ &= 0.38 \end{aligned}$$

3.2 If your friend forgot to water it, what's the chance it'll be dead when you return?

$$P(D|W') = 0.80$$

3.3 If it's dead when you return, what's the chance your friend forgot to water it?

$$\begin{aligned} P(W'|D) &= \frac{P(D|W')P(W')}{P(D)} \\ &= \frac{0.80 * 0.30}{0.38} \\ &= \frac{0.24}{0.38} \\ &= 0.6316 \end{aligned}$$

Question 4

4.1 Naive Bayes is a probabilistic model based on Bayes Theorem. Rewrite the formula below as G and B are conditionally independent given $D+$. Also, write about $P[D = -|G = g, B = b]$. How would you use the two formulas to determine diabetes when you have a glucose and blood pressure record (g, b)?

$$P[D = +|G = g, B = b] = \frac{P[G = g, B = b|D = +] \cdot P[D = +]}{P[G = g, B = b]}$$

We can rewrite the formula as:

$$P[D = +|G = g, B = b] = \frac{P[G = g|D = +] \cdot P[B = b|D = +] \cdot P[D = +]}{P[G = g] \cdot P[B = b]}$$

And the formula for $P[D = -|G = g, B = b]$ is:

$$P[D = -|G = g, B = b] = \frac{P[G = g|D = -] \cdot P[B = b|D = -] \cdot P[D = -]}{P[G = g] \cdot P[B = b]}$$

We can use both formulas to determine diabetes when we have a glucose and blood pressure record by comparing the probabilities of $P[D = +|G = g, B = b]$ and $P[D = -|G = g, B = b]$. If $P[D = +|G = g, B = b] > P[D = -|G = g, B = b]$, then the patient has diabetes, otherwise they don't.

4.4 Evaluate your classifier using "test.csv". Use accuracy rate
An Accuracy of 0.92 was found when evaluating the classifier using the test.csv file.

4.5 Do you think the standardization for data was necessary when building your Naive Bayes classifier? If yes, then why? If not, why we don't need to?
Standardization was not necessary when building the Naive Bayes classifier. This is because the Naive Bayes classifier is not affected by the scale of the data. The classifier only needs to know the probability of each feature given each class.

4.6 Do you think the data reflects reality well? Which part of the previous steps would you like to change if we cannot collect the data again? How would you change?
The data does not refect realtiy as there are other factors to the cause of diabetes. If we cannot collect the data again, one could do feature engineering to generate additional features.

Question 5

We have 10,000 3-D data points and computed mean and covariance information as below.

5.1 Let Y be a random vector defined by $\vec{Y} = A\vec{X} + \vec{b}$. Express $E[Y]$ and $COV[Y, Y]$ in terms of $E[X]$ and $COV[X, X]$.
For $E[Y]$, we have:

$$\begin{aligned} E[Y] &= E[A\vec{X} + \vec{b}] \\ &= E[A\vec{X}] + E[\vec{b}] \\ &= AE[\vec{X}] + \vec{b} \end{aligned}$$

For $COV[Y, Y]$, we have:

$$\begin{aligned} COV[\vec{Y}, \vec{Y}] &= E\left[(\vec{Y} - E[\vec{Y}])(\vec{Y} - E[\vec{Y}])^T\right] \\ &= E\left[(A\vec{X} - E[A\vec{X}])(A\vec{X} - E[A\vec{X}])^T\right] \\ &= E\left[(A\vec{X} - AE[\vec{X}])(A\vec{X} - AE[\vec{X}])^T\right] \\ &= E\left[A(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^T A^T\right] \\ &= AE\left[(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^T\right] A^T \\ &= ACOV[\vec{X}, \vec{X}]A^T \end{aligned}$$

5.2 Design A and b to whiten Y. i.e. E[Y] = 0 and COV [Y, Y] = I

To turn $E[Y] = 0$, we can set $AE[\vec{X}] + \vec{b} = 0$. Therefore

$$b = -AE[\vec{X}]$$

To turn $COV[Y, Y] = I$, we can set $ACOV[\vec{X}, \vec{X}]A^T = I$. To do this A must be the inverse square root of the covariance matrix of X We can calculate A by first finding the Eigen Decomposition of the covariance matrix of X, $COV[\vec{X}, \vec{X}] = U\Lambda U^T$.

After doing the eigen Decomposition, we the eigenvectors as:

$$\begin{bmatrix} 0.866 & -0.49 & 0 \\ 0.49 & 0.866 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

And the eigenvalues as:

$$\begin{bmatrix} 2.99 \\ 2 \\ 1 \end{bmatrix}$$

We can then use this eigen decomposition to calculate A as $A = U\Lambda^{-1/2}U^T$.
By calculating A we get:

$$\begin{bmatrix} 0.61 & 0.06 & 0 \\ -0.06 & 0.67 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We can then calculate b as $b = -AE[\vec{X}]$. After calculating b we get:

$$\begin{bmatrix} -0.105 \\ -0.191 \\ -0.1 \end{bmatrix}$$

Question 6

Suppose we want to measure a length, for example, the water depth µ at 79.137°(N) and 2.817°(E). The depth was measured repeatedly and recorded as $x_1, x_2, \cdots, x_n$. For device imperfection, the samples were varied by ϵ where $\epsilon \sim N(0, \sigma 2)$. i.e $x = \mu + \epsilon$. Bayes rule allows us to evaluate the uncertainty in μ after observing $\vec{x}$ in the posterior probability as below.

6.1 Given the observations $x_1, x_2, \cdots, x_n$ derive a formula to estimate μ^*_{ML} when μ is a fixed value. We assume the observations are i.i.d (independent and identically distributed) and follow multivariate Gaussian

The probablity density function for each observation x_i is:

$$P(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

And the log likelihood function is:

$$\begin{aligned}\log L(\mu, \sigma^2) &= \sum_{i=1}^{n} \log P(x_i|\mu, \sigma^2) \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \\ &= \sum_{i=1}^{n} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\end{aligned}$$

To find μ_{ML}^8 we can differentiate the log likelihood function with respect to μ and set it to 0:

$$\begin{aligned}\frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - n\mu \\ 0 &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - n\mu \\ n\mu &= \sum_{i=1}^{n} x_i \\ \mu_{ML}^* &= \frac{1}{n} \sum_{i=1}^{n} x_i\end{aligned}$$

6.2 Given the observations $x_1, x_2, \cdots, x_n$ derive a formula to estimate μ_{MAP}^* when μ is known to follow Gaussian $p(\mu) \sim N(\mu_0, \sigma_0^2)$

The random variable PDF is:

$$f(x_n|\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left(-\frac{(x_n - \mu)^2}{2\sigma_0^2} \right)$$

and the prior PDF is:

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right)$$

The MAP estimate is:

$$\mu_{MAP}^* = \arg\max_{\mu} \exp \left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right)$$

To find μ^*_{MAP} we can differentiate the log likelihood function with respect to μ and set it to 0:

$$\begin{aligned}
\frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu) - \frac{1}{\sigma_0^2}(\mu - \mu_0) \\
&= \frac{1}{\sigma_0^2} \sum_{i=1}^{n} x_i - n\mu - \frac{1}{\sigma_0^2}\mu + \frac{1}{\sigma_0^2}\mu_0 \\
0 &= \frac{1}{\sigma_0^2} \sum_{i=1}^{n} x_i - n\mu - \frac{1}{\sigma_0^2}\mu + \frac{1}{\sigma_0^2}\mu_0 \\
n\mu + \mu &= \sum_{i=1}^{n} x_i + \mu_0 \\
\mu^*_{MAP} &= \frac{1}{n+1} \sum_{i=1}^{n} x_i + \frac{1}{n+1}\mu_0
\end{aligned}$$

6.3 MAP vs ML

As N goes to infinity, the MAP estimate will converge to the ML estimate, as N goes to 0, the MAP estimate will converge to the prior mean. Use MLE when you have a lot of data and you are confident in the prior. Use MAP when you have a small amount of data and you want to incorporate the prior information.