National University of Singapore
School of Computing
CS2109S: Introduction to AI and Machine Learning
Semester I, 2023/2024

Tutorial 5
Classification and Logistic Regression

These questions will be discussed during the tutorial session. Please be prepared to answer them.

Summary of Key Concepts

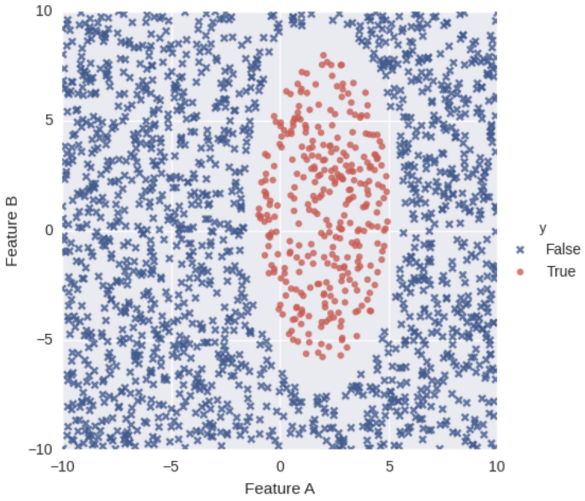
In this tutorial, we will discuss and explore the following learning points from Lecture:

1. Classification
   (a) Linear vs Non-linear Separability
2. Logistic Regression
   (a) Loss function
   (b) Choosing suitable metric
   (c) Multi-class classification
3. Performance measures
   (a) Confusion matrix
   (b) Precision vs Recall
   (c) ROC curve

Problems

1. **Linear vs Non-linear Separability**

   Bondrewd Workshop is a company that produces cute fluffy bunnies through experimentation and genetic mutations. Quality control for the bunnies is done manually. A group of scientists decide whether a bunny is ready to be released into the wild based on two features: **Feature A** is a bunny's cuteness score and **Feature B** is a bunny's fluffiness score. The figure below show examples of bunnies that have been released and withheld in the past. Each dot corresponds to a bunny, and responds to the following question as true or false: this bunny is ready to be released.



Knowing that you are a student in CS2109S, Bondrewd the CEO has approached you and asked you to automate the decision making process.

(a) Which set of features that will perfectly classify whether or not a bunny can be released into the wild. Here are some examples of sets of features:

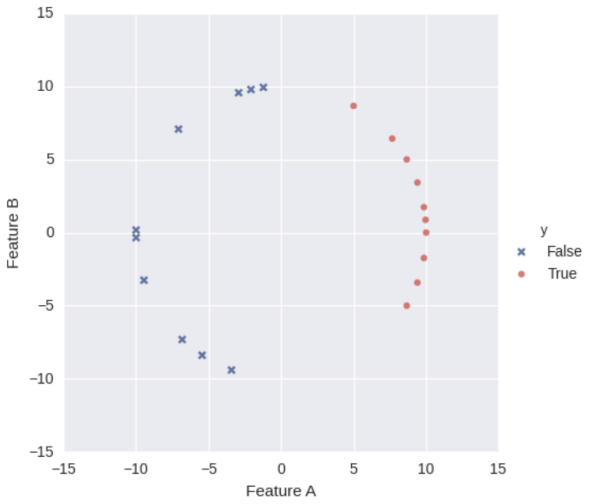
i. (A, B)

ii. (A)

iii. (B)

Solution: *Notice that an ellipse with major and minor axis parallel to y-axis and x-axis is sufficient to classify the data. Hence* (A^2, B^2, A, B) *minimally suffices. For more general ellipses (or conics) you can use the more general set of features:* (A^2, AB, B^2, A, B)*.*

(b) Bondrewd decides to change the production direction in the company. Bondrewed Workshop will be creating fewer, but cuter (and fluffier) bunnies. After more experiments, they have collected the examples again in the figure below.



Define a set of features that will perfectly classify whether or not a bunny can be released into the wild.

Solution: *We can use* (A) *and that is minimum set of features possible as a single vertical line is enough to classify. We can also use the set from 1(a) since a higher polynomial set of features can degenerate into a line.*

2. **Loss Function of Logistic Regression**.

In lecture, we discussed about Logistic Regression model which has the following hypothesis:

$$h_w(x) = \frac{1}{1 + e^{-w^T x}}$$

$h_w(x)$ could be interpreted as a probability p assigned by the model such that $y = 1$. The probability of $y = 0$ is therefore $1 - p$.

(a) Write down the probability p as a function of x and calculate the derivative of $\log(p)$ with respect to each weight w_i.

Solution:

First we write the probabilty p as a function of x.

$$p = \frac{1}{1 + e^{-w^T x}} = \frac{1}{1 + e^{-w \cdot x}} = \frac{1}{1 + e^{\sum_{i=1}^{n} -w_i x_i}}$$

Take the log of both sides,

$$\log(p) = \log\left(\frac{1}{1 + e^{\sum_{i=1}^{n} -w_i x_i}}\right) = -\log(1 + e^{\sum_{i=1}^{n} -w_i x_i})$$

Now we differentiate $\log(p)$ with respect to w_i

$$\begin{aligned}
\frac{\partial \log(p)}{\partial w_i} &= -\left(\frac{1}{1 + e^{\sum_{i=1}^{n} -w_i x_i}} \frac{\partial}{\partial w_i}(1 + e^{\sum_{i=1}^{n} -w_i x_i})\right) \\
&= -p\frac{\partial}{\partial w_i}(1 + e^{\sum_{i=1}^{n} -w_i x_i}) \\
&= -p(-x_i)e^{\sum_{i=1}^{n} -w_i x_i} \\
&= \boxed{(1-p)x_i}
\end{aligned}$$

(b) Write down the probability $1 - p$ as a function of x and calculate the derivative of $\log(1 - p)$ with respect to each weight w_i.

Solution:

First we write the probabilty $1 - p$ as a function of x.

$$1 - p = \frac{e^{-w^T x}}{1 + e^{-w^T x}} = \frac{1}{1 + e^{w^T x}} = \frac{1}{1 + e^{w \cdot x}} = \frac{1}{1 + e^{\sum_{i=1}^{n} w_i x_i}}$$

Take the log of both side,

$$\log(1 - p) = \log\left(\frac{1}{1 + e^{\sum_{i=1}^{n} w_i x_i}}\right) = -\log(1 + e^{\sum_{i=1}^{n} w_i x_i})$$

Now we differentiate $\log(1 - p)$ with respect to w_i

$$\begin{aligned}
\frac{\partial \log(1-p)}{\partial w_i} &= -\left(\frac{1}{1+e^{\sum_{i=1}^{n} w_i x_i}} \frac{\partial}{\partial w_i}(1+e^{\sum_{i=1}^{n} w_i x_i})\right) \\
&= -(1-p)\frac{\partial}{\partial w_i}(1+e^{\sum_{i=1}^{n} w_i x_i}) \\
&= -(1-p)(x_i)e^{\sum_{i=1}^{n} w_i x_i} \\
&= -(1-p)(x_i)\left(\frac{p}{1-p}\right) \\
&= \boxed{-px_i}
\end{aligned}$$

(c) Using results from 3(a) and 3(b), derive $\frac{\partial L}{\partial w_i}$, where L is the loss function of logistic regression model.

$$L = -y\log(h_w(x)) - (1-y)\log(1-h_w(x))$$

Solution:

First we substitute $h_w(x)$ as p

$$L = -y\log(p) - (1-y)\log(1-p)$$

Now we differentiate L with respect to w_i

$$\begin{aligned}
\frac{\partial L}{\partial w_i} &= -y\frac{\partial \log(p)}{\partial w_i} - (1-y)\frac{\partial \log(1-p)}{\partial w_i} \\
&= -y(1-p)x_i - (1-y)(-px_i) \\
&= -x_i(y-p) \\
&= \boxed{x_i(h_w(x)-y)}
\end{aligned}$$

3. **Evaluating Logistic Regression**.

Which of the following evaluation metrics is the least appropriate when comparing a logistic regression model's output with the target label? Explain your answer.

(a) Accuracy

(b) Binary Cross Entropy Loss

Binary Cross Entropy Loss is commonly used as an evaluation metric for logistic regression. It is also used as the cost function for logistic regression:

$$L = -ylogh_w(x) - (1-y)log(1-h_w(x))$$

Simply put, it is the negative log of the predicted probabilities.

(c) Mean Squared Error

Mean Squared Error (MSE) is commonly used as an evaluation metric for linear regression, but can be used for logistic regression as well. Here, it is defined as the squared difference between the predicted probability and the target label (0 if negative and 1 if positive, or vice versa) :

$$MSE = \frac{1}{2}(y - h_w(x))^2$$

(d) AUC-ROC

AUC-ROC curve tells how well the classifier is able to separate positive classes from negative classes. AUC refers to the area under a ROC curve. A model with an AUC score of 1 means it can separate positive and negative classes perfectly. On the other hand, a model with an AUC score of 0 means it is predicting the opposite class for each data (positive class is predicted as negative class and vice versa).

Solution: *Mean Squared Error*

For this question, it is important to distinguish between the evaluation metric, which assesses the performance of a model in a specific way, and the loss function, which is a mathematical expression used during the training of the model to guide parameter optimization. Students should understand the differences between those two, and discuss this question with respect to evaluation metrics.

Evaluation Metric for Classification

- *In classification tasks, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Binary Cross-Entropy (BCE) Loss are not suitable metrics for evaluating model performance as they may not capture the accuracy or appropriateness of class predictions.*
- *Accuracy is a useful metric for gauging a model's overall performance by indicating the proportion of correct predictions it made on the test set.*
- *The AUC-ROC curve is a valuable metric for assessing classifier performance in classification tasks as it quantifies the ability to distinguish between positive and negative classes effectively, making it the most informative measure in this context.*
- *Consider three data points X with labels $y = [0, 0, 1]$ and two logistic regression models, M_1 and M_2. M_1 predicts the three data points $\hat{y} = [0.4, 0.4, 0.6]$, while M_2 predicts the three data points as $\hat{y} = [0.1, 0.6, 0.9]$. If we use **MSE**, M_1 will have a loss of $\frac{2}{25} = 0.08$, while M_2 will have a loss of $\frac{19}{300} \approx 0.063$. The loss*

of $M_1 > M_2$, but M_1 predicts more data points correctly than M_2. This example can be extended to MAE and BCE Loss too.

Loss Function

- When assessing the proximity of predicted probabilities to the actual values, Mean Squared Error (MSE) and Mean Absolute Error (MAE) can be considered, but Binary Cross-Entropy (BCE) Loss is often more appropriate for classification tasks.
- It's important to note that BCE Loss may not be suitable when dealing with cases where a single data point has all incorrect predictions, potentially resulting in infinite loss.

Bonus: What about Mean Absolute Error (MAE)?

Similar to MSE, Mean Absolute Error (MAE) is defined here as:

$$MAE = \frac{1}{2}|y - h_w(x)|$$

MAE suffers from the same issue as **MSE**. Consider three data points X with labels $y = [0, 0, 1]$ and two logistic regression models, M_1 and M_2. M_1 predicts the three data points $\hat{y} = [0.2, 0.4, 0.6]$, while M_2 predicts the three data points as $\hat{y} = [0.1, 0.6, 0.9]$. If we use **MAE**, M_1 will have a loss of $\frac{1}{3} \approx 0.333$, while M_2 will have a loss of $\frac{4}{15} \approx 0.267$. The loss of $M_1 > M_2$, but M_1 predicts more data points correctly than M_2.

4. **Logistic Regression for Multi-Class Classification**.

Suppose you have a classification task of deciding whether an animal is a cat, a horse, or an elephant. However, you can't see the animal but you have the information about

- The weight of the animal (in kilogram)
- The length of the animal (in meter)

You, being an ML Expert, suggested to use 3 Logistic Regression models to solve this problem. After training on the training dataset, you get the following parameters:

$$w_{cat} = [4.2, -0.01, -0.12]$$
$$w_{horse} = [-20, -0.08, 35]$$
$$w_{elephant} = [-1250, 0.82, 0.9]$$

(a) You're given a list of animals with their features. Compute the probability of an animal belonging to a certain class and classify them accordingly.

Weight (kg)	Length (m)
4.2	0.4
720	2.4
2350	5.5

Table 1: List of animals with unknown class

Solution:

For the first animal,

$$p_{cat} \approx 1 \quad p_{horse} = 0.00177 \quad p_{elephant} \approx 0$$

Hence, we classify the first animal as a cat.

For the second animal,

$$p_{cat} = 0.036 \quad p_{horse} = 0.999 \quad p_{elephant} \approx 0$$

Hence, we classify the second animal as a horse.

For the third animal,

$$p_{cat} \approx 0 \quad p_{horse} \approx 0 \quad p_{elephant} \approx 1$$

Hence, we classify the third animal as an elephant.

(b) What if we want to extend the classification task to classify other animals? Can we train a new model while keeping the weights of the previous models?

Solution: *It depends. For an animal that are very distinct with the three animals, we can create a new logistic regression model without changing the previous weights. However, for classifying a new animal that is similar with one of the classes (e.g, classifying a dog), we need to retrain the old models.*

5. **Precision, recall, F1 score and ROC curve**

Esophageal cancer is a serious and very aggressive disease. In this question, we want to look at the size of a patient's tumor and decide whether the cancer has spread to his or her lymph nodes. Using what we learnt, we use maximum dimension (mm) of esophagus tumor as input, and label 1 if the cancer had spread to their lymph nodes, and 0 otherwise. We derived this machine learning model M which outputs a continuous score for every input sample. Figure 1 shows the results for 20 samples from the model. The actual labels can be either 1 (red, positive label) or 0 (blue, negative label). The model output makes the final classification decision. If a threshold, p is given, model M outputs label 1 if M(x) is greater than or equal to the threshold, otherwise the model outputs 0.

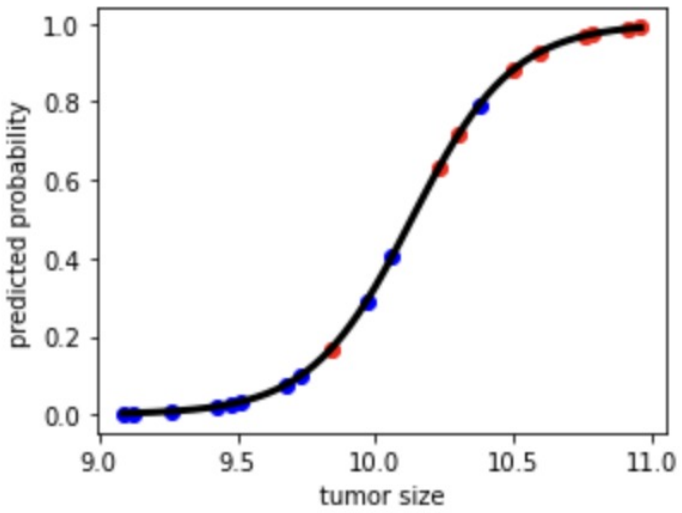


Figure 1: Model probability output and tumor size

(a) For the threshold p = 0.5, come up with the confusion matrix.

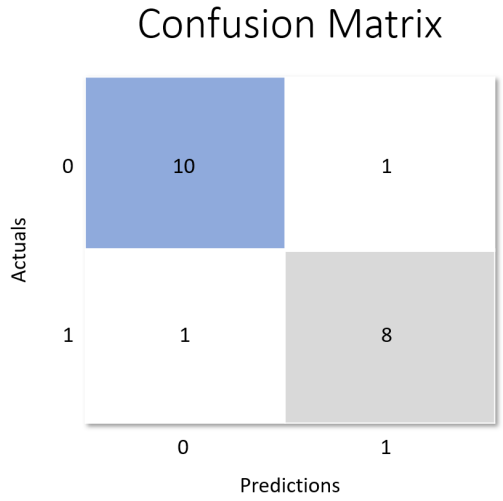


Figure 2: Confusion matrix

(b) For the threshold p = 0.5, find the precision, recall and F1 score.

Hint: Precision = $\frac{TP}{TP+FP}$ *Recall* = $\frac{TP}{TP+FN}$ *F1 score* = $\frac{2TP}{2TP+FP+FN}$

Solution: *We find that TP = 8, FP = 1, TN = 10, FN = 1. Precision, Recall, and F1 scores can be calculated as follows.*

$$Precision = \frac{TP}{TP+FP} = \frac{8}{8+1} = \frac{8}{9}$$

$$Recall = \frac{TP}{TP+FN} = \frac{8}{8+1} = \frac{8}{9}$$

$$F1 = \frac{2*TP}{2*TP+FP+FN} = \frac{2*8}{2*8+1+1} = \frac{8}{9}$$

(c) Based on figure 1, derive the ROC curve.

Hint: TPR = $\frac{TP}{ActualPositive}$ = $\frac{TP}{TP+FN}$ *FPR* = $\frac{FP}{ActualNegative}$ =$\frac{FP}{TN+FP}$

Hint 2: Tabulate a confusion matrix and from there, calculate the true positive rates and false positive rates. Mark out the corresponding point on the graph. Repeat this for at least four different thresholds.

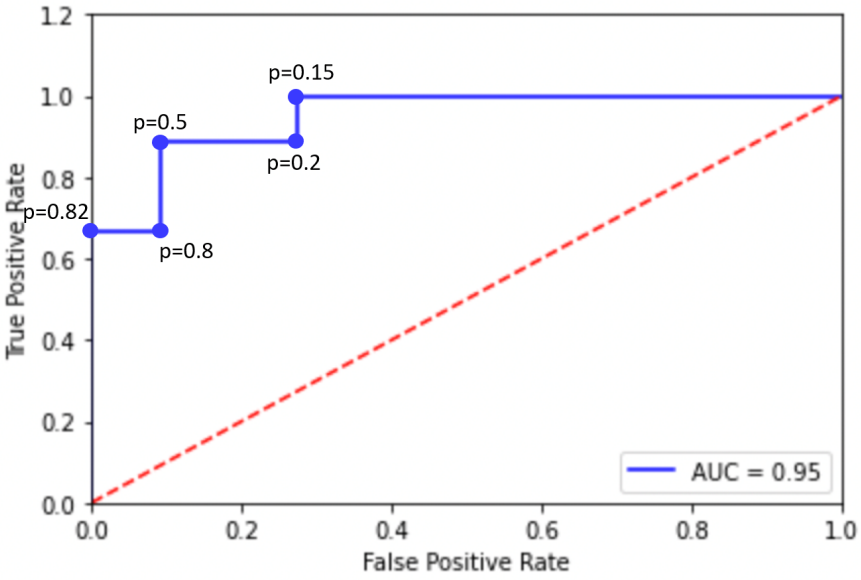


Figure 3: ROC curve

(d) **(Optional)** Based on the ROC curve you derived, decide which threshold you want to choose among p = 0.2, p = 0.5 and p = 0.8.

Solution: *Among these three thresholds, we should choose p = 0.5. p = 0.2 and p = 0.5 gives the same true positive rate, but false positive rate for p = 0.2 is higher. p = 0.5 and p = 0.8 gives the same false positive rate, but true positive rate for p = 0.5 is higher. Hence we choose p = 0.5.*

(e) **(Optional)** In this question's case for detecting tumours, should we maximize precision or recall? Explain the reason for your choice.

Solution: *If cancer detection is being performed as a regular check up, then precision should be maximized; as we do not want to start cancer treatment on a person unless we are sure that he has cancer. On the other hand, if cancer detection is being performed as part of cancer treatment progress monitoring, then recall should be maximized; as we do not want to stop the ongoing treatment unless we are sure that there is no residual tumour cell left in the patient.*

(f) **(Optional)** Suppose now we want to detect plagiarism instead, should we maximize precision or recall? Explain the reason for your choice.

Solution: *In this case, we should maximize precision. This is because we don't want to wrongly accuse those who did not plagiarize. Therefore, we should minimize false positives.*

Bonus: Recall that we were helping banks with credit card fraud detection in PS5. In such a case, should we maximize precision or recall? Explain the reason for your choice.

Solution: *In this case, there isn't a clear guideline on which to maximize. If we allow a lot of false positives (if fraud is the positive label here), we will incorrectly think that some transactions are fraudulent and block them, causing inconvenience to the customers. On the other hand, if we allow for a lot of false negatives, then we will miss a lot of fraudulent cases, and the bank will lose a lot of money. Given these, in this case, we should strike a balance between recall and precision.*