CS2109s Tutorial 5

by Lee Zong Xun

How was your midterms?

- If it went well, congratulations!
- If it didn't, don't worry, you can still do well in the final exam!
- If you have any questions, feel free to ask me!

Recap

What is **classification**?

- Given a set of data points, we want to classify them into different categories.

How to draw a **decision boundary**?

- If data **is linearly separable**, just draw a decision boundary that separates the data (a line!).
- If the data is **not linearly separable**, we can use a **kernel trick** to transform the data into a higher dimension, and then draw a decision boundary in the higher dimension. (Will discuss more in the next tutorial)

Recap

What's wrong with simply using the threshold function $h(x) = 1$ if $w^T x + b > 0$ and $h(x) = 0$ otherwise?

- The function is not **differentiable** and not **continuous**. We cannot iteratively find the optimal weights and bias!
- We introduce the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ to solve this problem, where $z = w^T x + b$. **Interpretation:** The probability that $y = 1$ on input x.

Can we use the same cost function as Linear Regression?

- No, because the cost function is not convex! We cannot find the global minimum using gradient descent.

Recap

What is the **cost function** for Logistic Regression?

- The cost function is the **cross-entropy** function, which is defined as

$$J(w,b) = -\frac{1}{m}\sum\_i = 1^m y^{(i)} \log(h_w(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_w(x^{(i)}))$$

- The gradient descent algorithm for Logistic Regression is the same as Linear Regression! We will show in this tutorial.

  The formula for the gradient descent algorithm is $w := w - \alpha \frac{\partial J(w,b)}{\partial w}$.

Recap

How do we pick the best hypothesis?

- Split the dataset into a **training set**, **validation set** and a **test set**.
- Train the model on the training set, and evaluate the model on the validation set.
- Pick the model with lowest validation error.
- Use the test set to evaluate the final model (for unseen samples).

Bias vs Variance

- A model with **high bias** makes more assumptions about the data, and is less flexible. It is more likely to underfit the data.
- A model with **high variance** makes fewer assumptions about the data, and is more flexible. It is more likely to overfit the data.

Recap

How to reduce overfitting?

- Reduce the magnitude of weights by adding an additional term!

  Cost function for **logistic regression** with **regularization**:

  $$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_w(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_w(x^{(i)})) + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

- Increasing λ will reduce the magnitude of the weights, and thus reduce overfitting. But it will also increase the bias.

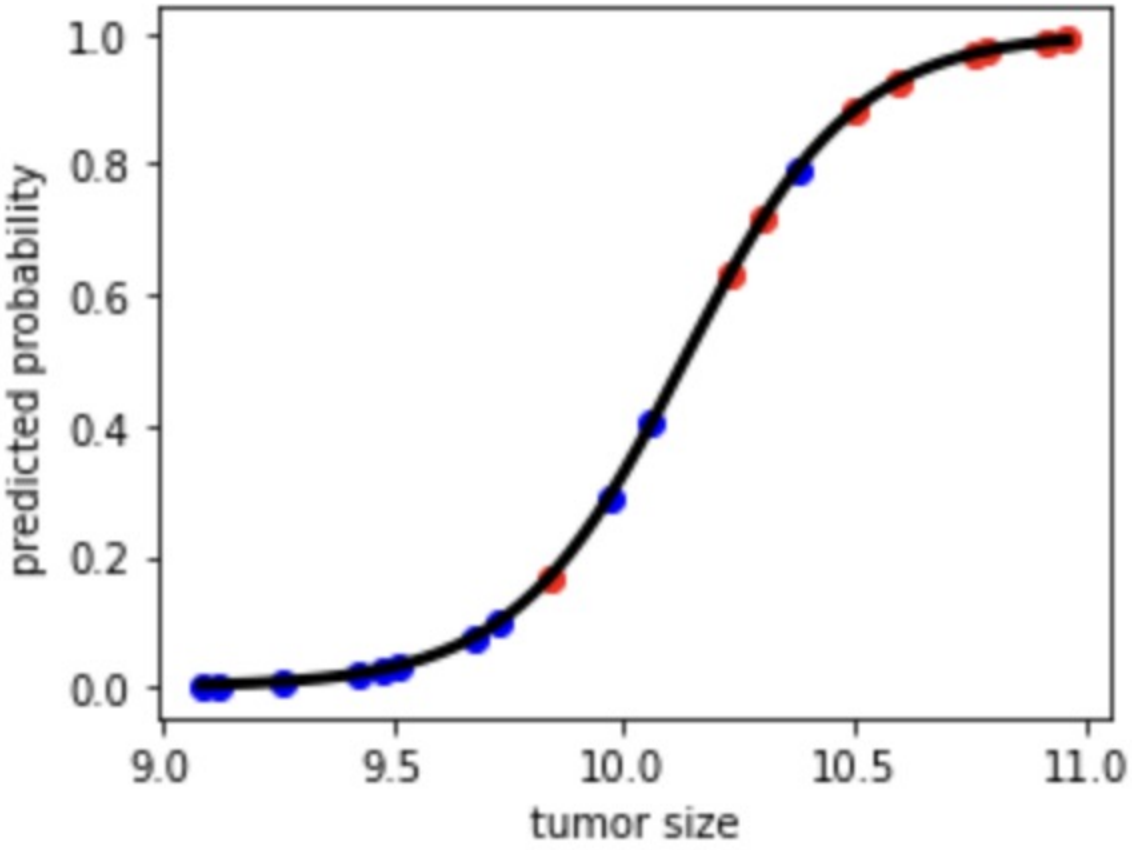
Any Questions?

Question 1

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.

Label 1 if the cancer had spread to their lymph nodes, and 0 otherwise. 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.

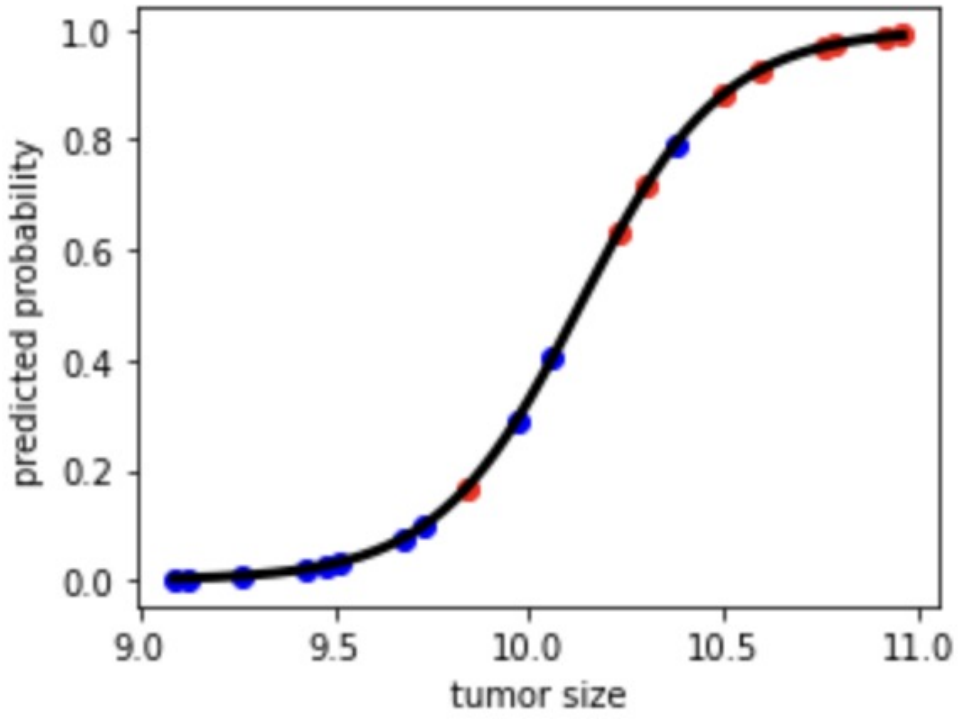
Question 1

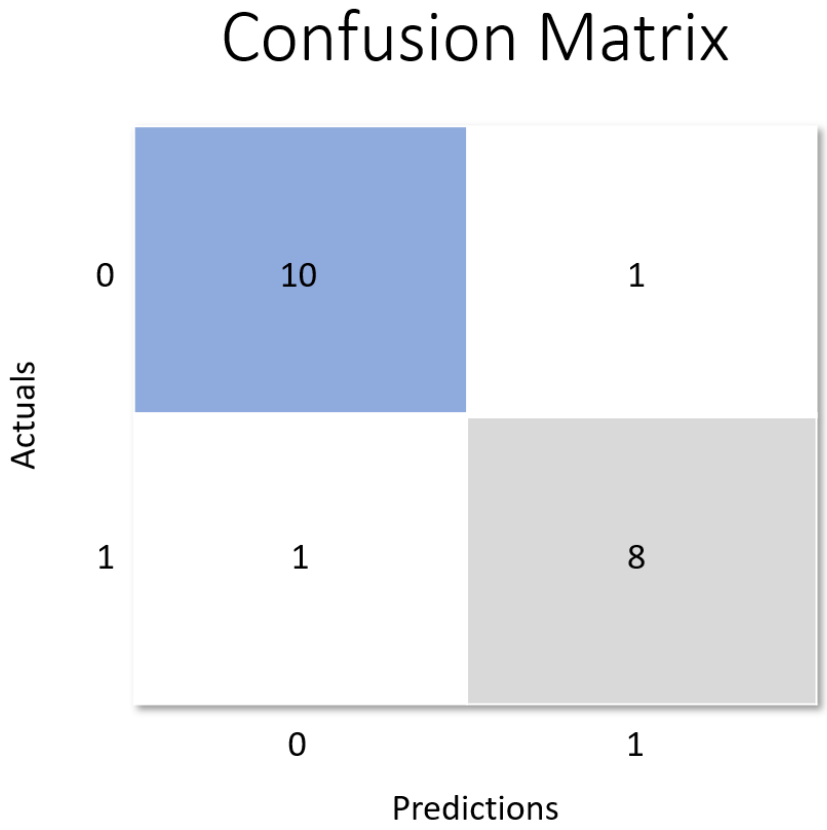


The actual labels can be either 1 (red, positive label) or 0 (blue, negative label).

Question 1(a)

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





Question 1(b)

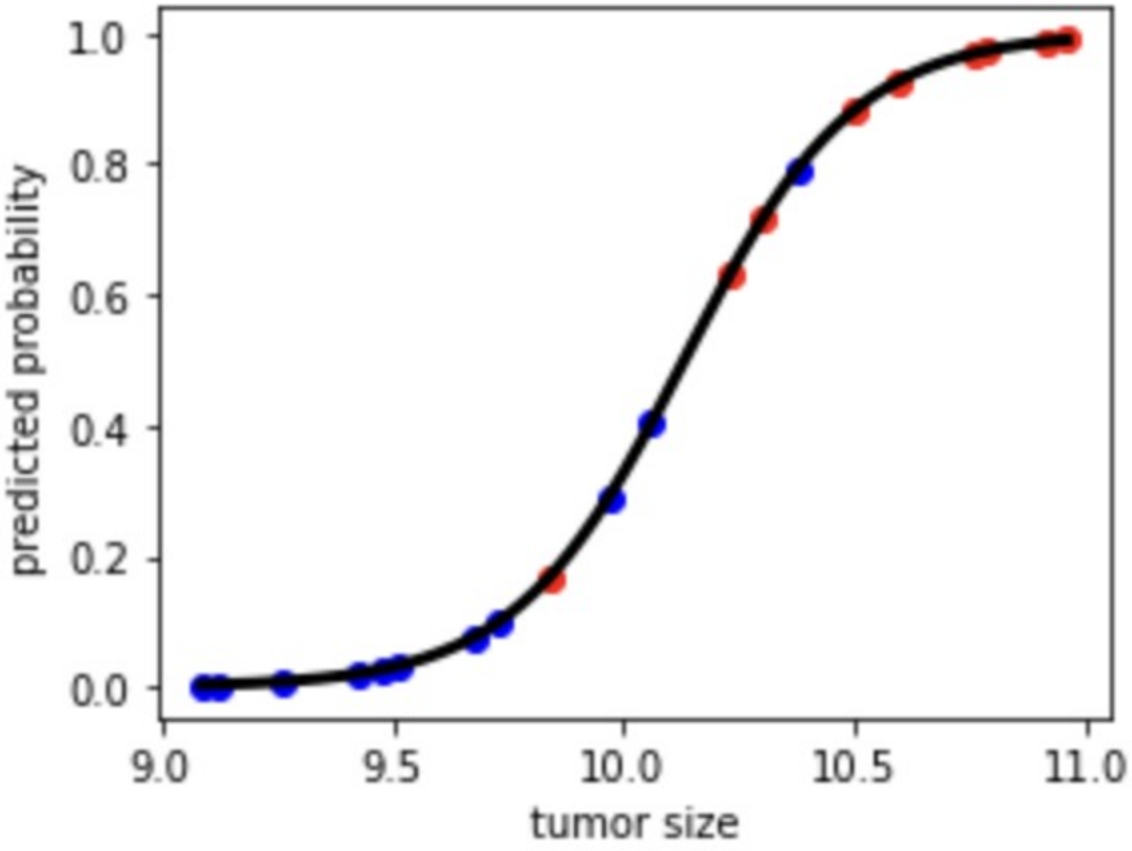
For the threshold p = 0.5, find the precision, recall and F1 score.

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

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

Question 1(c)

Based on figure 1, derive the ROC curve.



What happens if we set the threshold to be $p = 0$?

What happens if we set the threshold to be $p = 0.1$?

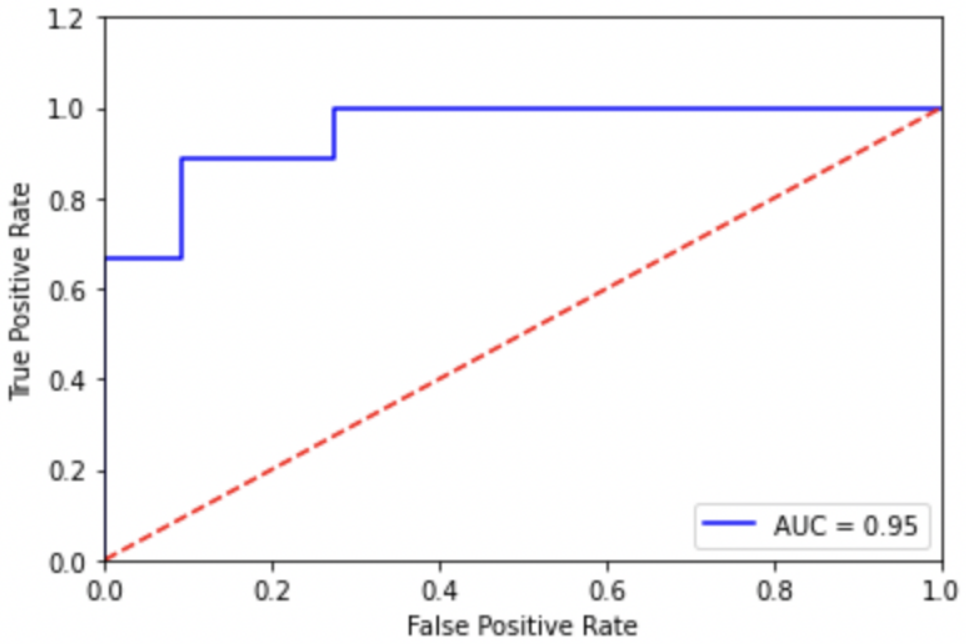
:

What happens if we set the threshold to be $p = 1$?

Question 1(c)

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

We derive the following plot...



$Threshold$	TPR	FPR
0	1	1
0.2	$8/9$	$3/11$
0.4	$8/9$	$2/11$
0.5	$8/9$	$1/11$
0.8	$6/9$	$1/11$
1	0	0

Question 1(d)

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

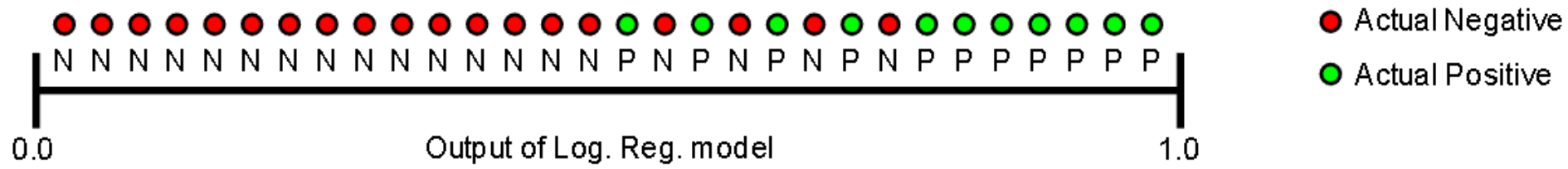
Among these three thresholds, we should choose p = 0.5.

- 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.

Before we move on...

What is the exact interpretation of the ROC curve?



- What does it mean if the AUC is equals to 1?
- What does it mean if the AUC is equals to 0?

Quick check

How would multiplying all of the predictions from a given model by 2.0 (for example, if the model predicts 0.4, we multiply by 2.0 to get a prediction of 0.8) change the model's performance as measured by AUC?

Answer

No change. AUC only cares about relative prediction scores. AUC is based on the relative predictions, so any transformation of the predictions that preserves the relative ranking has no effect on AUC.

Optional: Is this also the case for other metrics such as squared error, log loss, or prediction bias? 🤔

Question 1(e)

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

- Do we want to minimize false positives or minimize false negatives? Why?
  - Maximize precision if false positives is costly
  - Maximize recall if false negatives is costly
- At the end of the day, what is the goal of the model?
- As usual, it depends on the context!

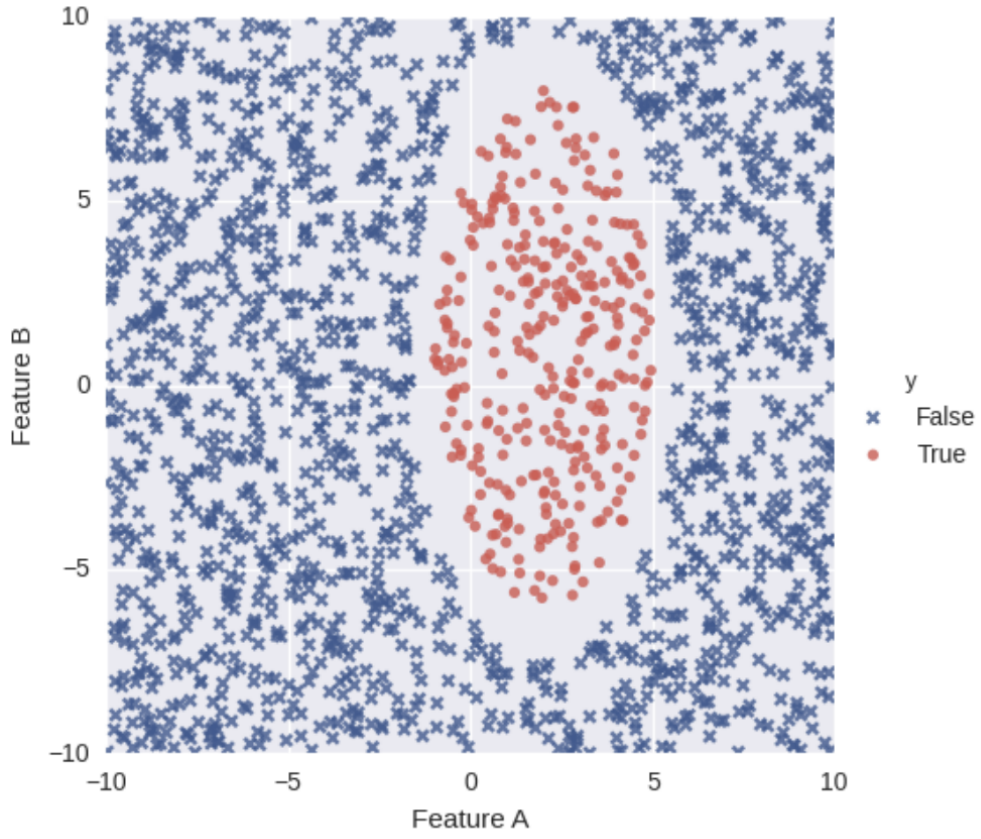
Question 1(f)

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

In this case, we should maximize precision. This is because we don't want to wrongly accuse those who did not plagiarize.

Question 2

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**
- **Feature B is a bunny's fluffiness score**

Question 2(a)

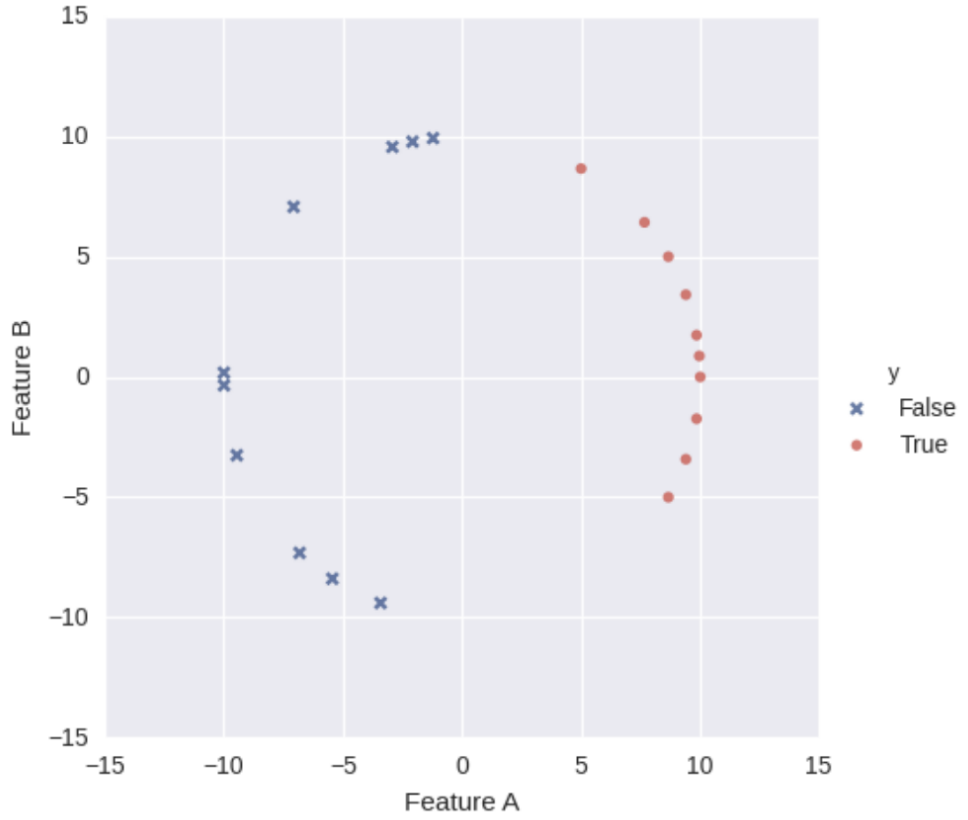
Define a reasonable set of features that will perfectly classify whether or not a bunny can be released into the wild.

Notice that an ellipse with major and minor axis parallel to y-axis and x-axis is sufficient to classify the data.

- (A^2, AB, B^2, A, B).

Question 2(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.



Question 2(b)

Notice an interesting pattern in the data.

- The data is now linearly separable. We can simply use feature A to classify the data!
- Can we use the same model from 2(a)? Yes! A higher polynomial set of features can degenerate into a line.

Question 3

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 $1 - p$.

Question 3(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.

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 side,

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

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

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

Question 3(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.

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

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

Question 3(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.

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

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

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

This is exactly the gradient descent update rule for logistic regression!

Question 4

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.

- Accuracy
- Log Loss
- Mean Squared Error
- AUC-ROC

- Accuracy is a useful metric that shows us how many predictions the model got right from the test set.
- Log loss is a good measure of how well the learning algorithm is doing.
- 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 \textbf{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 closer than M_2.
- The AUC-ROC curve tells how well the classifier is able to separate positive classes from negative classes.

Quick check

What is the difference between evaluation metrics and loss functions?

Evaluation metrics for a model explain the performance of a model

Loss functions are used to optimize your model during training by minimizing the loss function using gradient-based training methods.

Optional: We know that AUC-ROC is a useful metric to understand the performance of classifier on separating positive classes from negative classes, but is it suitable as a loss function?

Question 5

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:

$$w\_cat = [4.2, -0.01, -0.12]$$

$$w\_horse = [0, -1, 3]$$

$$w\_elephant = [-1250, 0.82, 0.9]$$

Question 5(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
320	160
2350	450

For the first animal,

$$p * cat = 0.984 \quad p * horse = 0.047 \quad p\_elephant \approx 0$$

Hence, we classify the first animal as a cat.

For the second animal,

$$p * cat = 1.25 \times 10^{-8} \quad p * horse \approx 1 \quad p\_elephant \approx 0$$

Hence, we classify the second animal as a horse.

For the third animal,

$$p * cat = 1.47 \times 10^{-32} \quad p * horse \approx 0 \quad p\_elephant \approx 1$$

Hence, we classify the third animal as an elephant (it's a **450 metres long** elephant 🤯)

Question 5(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?

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.
- 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.