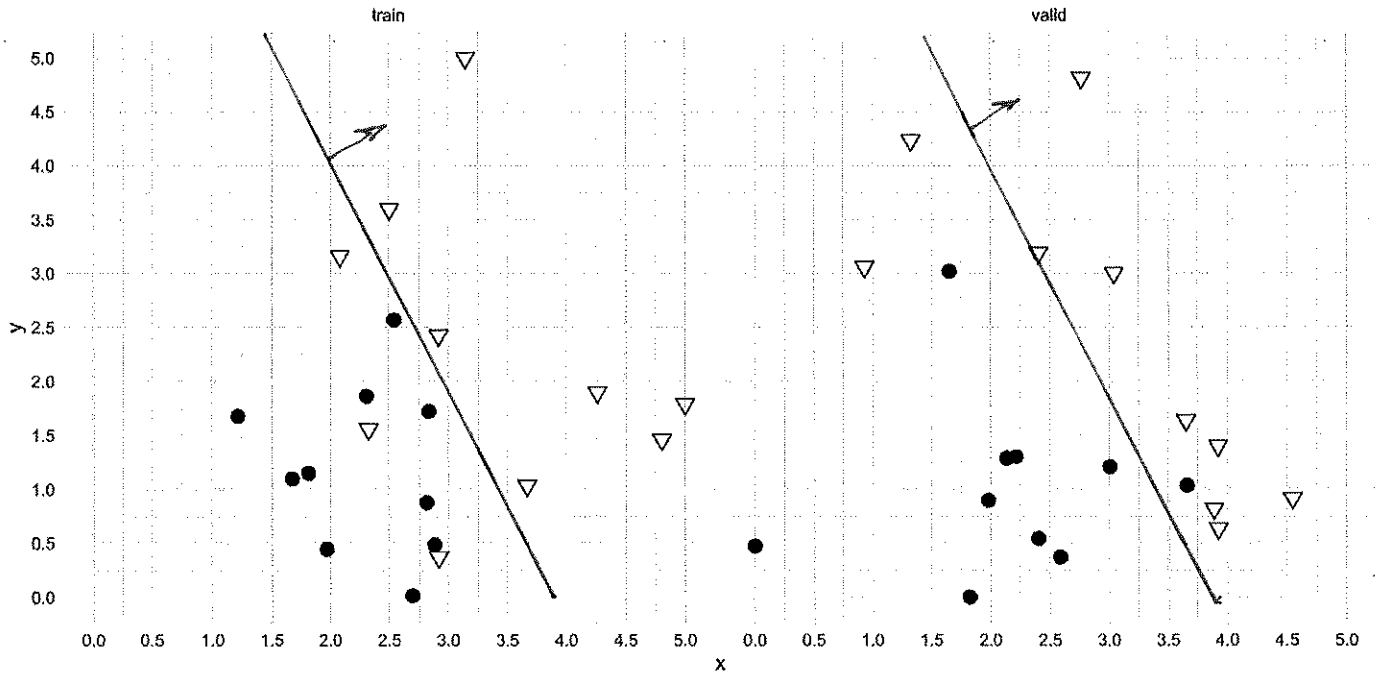


1. Above is some data similar to today's slides, expect that shape has used to describe the classes rather than the blue/orange. Start by covering up the validation data (right) and then answer the questions below.

a. Draw freehand an estimate of a good separating line for the points in the training data.
b. Now, uncover the validation data and create the same line from (a) on the right.
c. Next, compute a confusion matrix for the validation data in the space below:

Actual \ Predicted	●	▽
●	9	1
▽	2	8

[your line and results may be different]

d. Calculate the error rate for the training data and the validation data.

$$\text{training} = \frac{3}{20} = .15$$

$$\text{validation} = \frac{3}{20} = .15$$

[features are just possible options; signs in parenthesis]

2. Five examples of predictive modelling tasks are described. For each, identify (i) the classes, (ii) what an observation is, (iii) two possible numeric features, and (iv) pick one class to be the reference class (the class you are trying to predict) and estimate the signs you would expect to be associated with each of the features.

a. Create a model to predict whether a basketball player makes a free-throw shot.

Classes: made, missed

observation: shot

features: distance to basket (+)

player height (+)

b. Predict whether it will rain tomorrow.

[could be neg]

Classes: rain, dry

observation: day

features: mm of rain this year (+)

mm of rain last year (+)

c. Estimate whether a user's cell phone will still be working one year from now.

Classes: works, broken

observation: cell phone or user

features: current cell age (-)

current cell price (+)

d. Predict whether a tweet is associated with positive or negative feelings.

classes: positive, negative

observation: one tweet

feature: # smiling emoji (+)

frowning emoji (-)

e. Predict whether a music song is classified as "rock" or "classical" music.

Classes: classical, rock

observation: song

feature: length of song (+)

loudness of song (-)

f. Create a model to determine if an image was taken indoors or outside.

classes: indoor, outdoor

observation: image

feature: brightness (-)

people (+)

3. Assume that we have built a logistic regression model to predict whether a text message is spam using two features: the count of the number of numeric digits 0-9 (we will call this X) and the count of the number of words with two or more letters that are written all in capitals (we will call this Y). Our training data yielded the following model:

$$\textbf{Prob(spam)} = F(-1.2 + 0.3 \times X + 0.2 \times Y)$$

Answer the following questions based on this model.

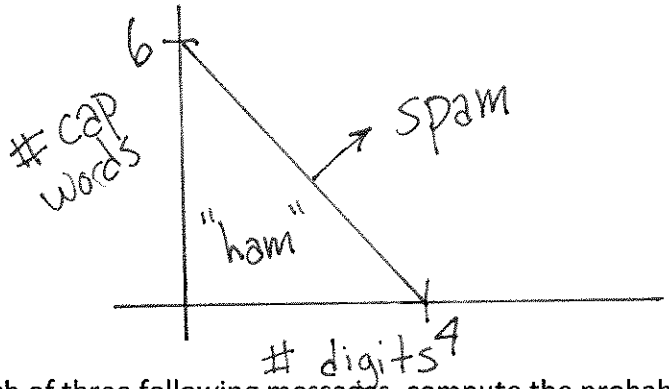
a. If a message contains no words with two or more letters that are written all in capitals, after how many numerical digits is there a predicted probability of at least 0.5 that it is spam?

Need $0.3X = 1.2 \Rightarrow$ 4 digits needed

b. If a message contains no numerical digits, after how many words with two or more letters that are written all in capitals is there a predicted probability of at least 0.5 that it is spam?

Need $0.2Y = 1.2 \Rightarrow$ 6 words needed

c. Based on (a) and (b), draw a sketch of the boundary line implied by the logistic regression model.



d. For each of three following messages, compute the probability that the text message is spam.

i. "Congratulations, you have won €12.000. Text us back ASAP to claim!"
ii. "Hi Joe. I owe you £30.50 for dinner last night. Do you have Venmo?"
iii. "I am taking MATH389 this semester. It is by far the WORST class I have ever taken."

i. $F(-1.2 + 0.3 \times 5 + 0.2 \times 1) = F(0.5) = 0.623$

ii. $F(-1.2 + 0.3 \times 4 + 0.2 \times 0) = F(0) = 1/2$

iii. $F(-1.2 + 0.3 \times 3 + 0.2 \times 1) = F(-0.1) = 0.475$

could be 2 if you count MATH389

4. I promised we would not do much in the way of formal mathematical calculations, but this one is relatively easy and conceptually important. We defined the logistic function F(x) = exp(x) / (exp(x) + 1). Use this to answer the two questions below.

a. Show that 1 – F(X) is equal to F(-X) for any number X.

$$1 - F(X) = 1 - \frac{\exp(X)}{\exp(X)+1} = \frac{\exp(X)+1}{\exp(X)+1} - \frac{\exp(X)}{\exp(X)+1}$$

$$= \frac{1}{\exp(X)+1} = \frac{1}{\exp(X)+1} \cdot \frac{\exp(-X)}{\exp(-X)} = \frac{\exp(-X)}{1+\exp(-X)} = F(-X)$$

b. From today's notes, assume we have built a two-feature logistic regression model **a + b × X + c × Y** to predict the probability that a point is orange. Use the result in (a) to calculate the logistic regression model to predict that probability that a point is blue. From the result, you should see that it is not particularly important which class we set as the reference class.

If

$$\text{Probability [orange]} = F(a + b \times X + c \times Y)$$

Then

$$\text{Probability [blue]} = 1 - \text{Probability [orange]}$$

$$= 1 - F(a + b \times X + c \times Y)$$

$$\text{from (a)} \rightarrow = F(-a - b \times X - c \times Y)$$

So same model for probability orange but with signs of parameters flipped.