Quiz #1 – Solutions

1. Suppose you are working on weather prediction and use a learning algorithm to predict tomorrow's temperature (in degrees Celsius/Fahrenheit). Would you treat this as a classification or a regression problem?

a) Regression
b) Classification

Solution: A

This question checked understanding of the distinction between regression (for a quantitative response variable) and classification (for a qualitative response variable).

2. Some of the problems below are best addressed using a supervised learning algorithm and the others with an unsupervised learning algorithm. Assuming that an appropriate dataset is available for your algorithm to learn from, which of the following could you apply supervised learning given only what is provided? (Select all that apply.)

a) In farming, given data on crop yields over the last 50 years, predict next year's crop yields.
b) Given historical data of children' ages and heights, predict children's height as a function of their age.
c) Examine a large collection of emails that are known to be spam email to discover if there are sub-types of spam mail.
d) Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.

Solution: A, B

This question checked understanding of the distinction between supervised learning (where you're trying to predict a response variable) and unsupervised learning (where you're trying to find more general patterns in a dataset). The first two problems described regression problems (hence supervised learning), while the last two described clustering applications (unsupervised).

3. Which of the following are true statements? (Select all that apply.)

a) Machine learning research has been developed primarily in the fields of computer science, engineering, and statistics.
b) Modern, computer-intensive, machine learning techniques are nearly always better than old, traditional, statistical methods.
c) Business decisions that are informed by data analysis tend to be better than those that are uninformed.

d) Many businesses successfully employ machine learning techniques to learn from customer behavior that they have stored in data warehouses.

Solution: A, C, D

All but one of the statements were true. Note that in many problems with a quantitative response variable it can be tough to beat good old multiple linear regression, while in many classification problems logistic regression often does very well. Message: when given a box of shiny new toys to play with, don't forget that the old toys can be just as much fun, sometimes more so.

4. Suppose we're trying to build a classification model to predict whether email is "spam" (junk email) based on information from a large number of emails. The objective is to design an automatic spam detector that could filter out spam before clogging a user's inbox. Ideally, we want to avoid filtering out good email, while letting spam get through is not desirable but less serious in its consequences. The classification model produces a prediction probability of an email being spam. Of the following choices, which best sets the “cut-off probability” (i.e., email with a prediction probability above the cutoff will be classified as spam)?

a) 0
b) Between 0 and 0.5
c) Between 0.5 and 1
d) 1

Solution: C

If false positives (predicting spam when it's not spam) are more costly than false negatives (predicting non-spam when it is spam), then the cut-off probability for predicting spam should be set higher than 0.5 to make the threshold for being classified as spam more difficult to attain. This will result in fewer emails being classified as spam, thus reducing the false positive rate. At the same time the false negative rate will increase but the net cost will decrease initially (since false positives are more costly than false negatives). As the cut-off probability continues to increase the net cost will be minimized at some probability between 0.5 and 1 and will then start to increase.

5. Why are machine learning techniques increasingly being used in business?

a) Large amounts of data are being produced and warehoused.
b) Computing power has increased over time and is widely affordable.
c) Interest in customer relationship management has increased as many companies move from a product-based to a service-based focus.
d) There are many machine learning software packages and algorithms available.
e) All of the above.

Solution: E

This question considered why machine learning has been gaining in popularity in recent years and all four reasons are relevant.

6. Consider the Wage Data described on pages 1 and 2 of the textbook. Suppose we developed a model to predict Wage solely using the trends visible in Figure 1.1 based on an employee's Age, Year, and Education Level. What is the most likely order for the predicted wages for the following individuals (from highest to lowest)?

a) Age 70, Year 2006, Education Level 3
b) Age 50, Year 2008, Education Level 4
c) Age 30, Year 2004, Education Level 4

Solution: Order = B (Highest), A (Second Highest), C (Lowest)

One way to answer this question was to estimate mean wages for each employee based on the graphs in Figure 1.1, which leads to a clear ordering for their overall predicted wages. For example, the "age 50, year 2008, education level 4" employee has the highest estimated mean wage of the three employees for all three graphs and so has the highest predicted wage overall.

7. Consider the Gene Expression Data described on pages 4 and 5 in the textbook, in particular the four distinct clusters of cell lines in the left-hand panel of Figure 1.4. If we just consider the first principal component, Z1, cell lines in the "red" cluster tend to have low values of Z1, cell lines in the "green" cluster tend to have high values of Z1, and cell lines in the "purple" and "blue" clusters tend to have medium values of Z1. If we also consider the second principal component, Z2, we can also distinguish between the "purple" and "blue" clusters because the former tend to have high values of Z2 while the latter tend to have low values of Z2. Suppose we also considered the third principal component, Z3, in a three-dimensional scatterplot. Say whether the following statement is true or false: It is possible that the three-dimensional scatterplot might reveal five or more distinct clusters if, for example, some of the cell lines in the "purple" cluster tended to have low values of Z3, while other cell lines in the "purple" cluster tended to have high values of Z3.

a) True
b) False

Solution: True

This question was essentially describing in practical terms how adding a predictor can reveal more detailed structure in a clustering application.

8. Which of the following are not supervised machine learning techniques? (Select all that apply.)

a) Logistic regression
b) Nearest neighbors
c) Classification and regression trees
d) Linear regression
e) Clustering

Solution: E

Clustering is the only unsupervised machine learning technique in this list.

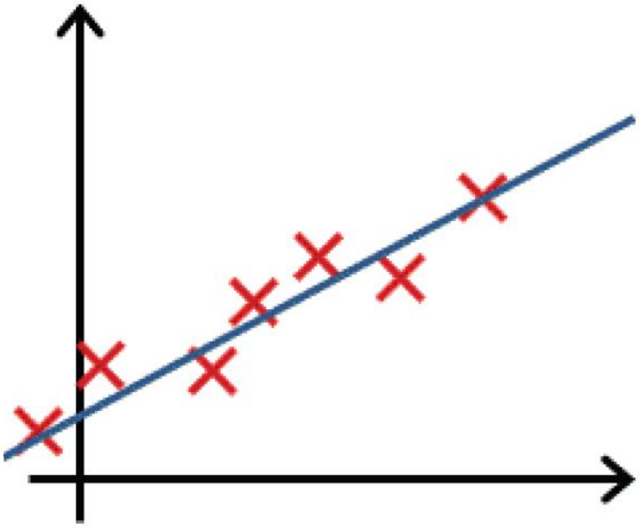
9. Suppose that it costs $1 to mail an offer, expected revenue if someone responds is $60, and the fixed costs of a mailing campaign are $20,000. A machine learning model orders prospective customers from most to least likely to respond so that mailing to 10,000 prospects yields a 6% response rate, mailing to 20,000 prospects yields a 4% response rate, and mailing to 30,000 prospects yields a 3% response rate. What are the net profits for mailings of size 10,000, 20,000, and 30,000? Write your answers in thousands of dollars (but don't include "000" or a dollar sign). For example, the net profit for 10,000 mailings is $6,000 so the correct answer is simply 6.

Solution:

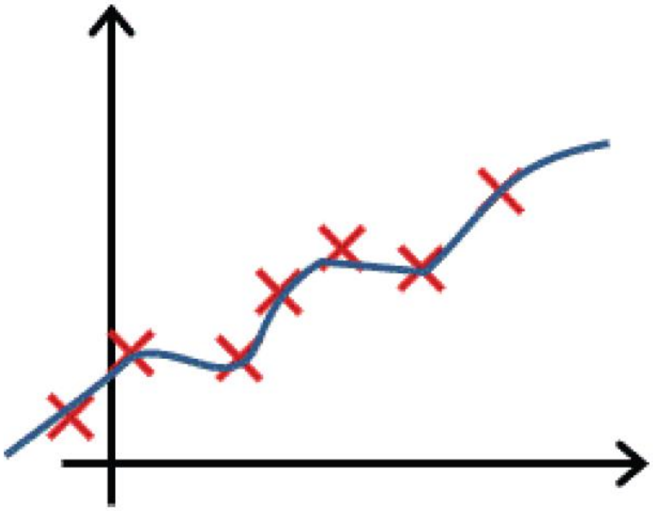
- 10,000 mailings = 6
- 20,000 mailings = 8
- 30,000 mailings = 4

Net profit = (# mailings * response rate * 60) - (# mailings * 1) - 20,000

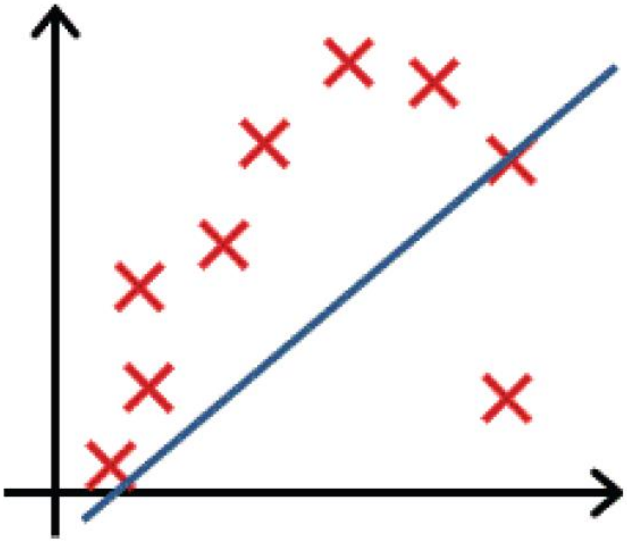
10. In which one of the following figures do you think the model has overfit the training set?



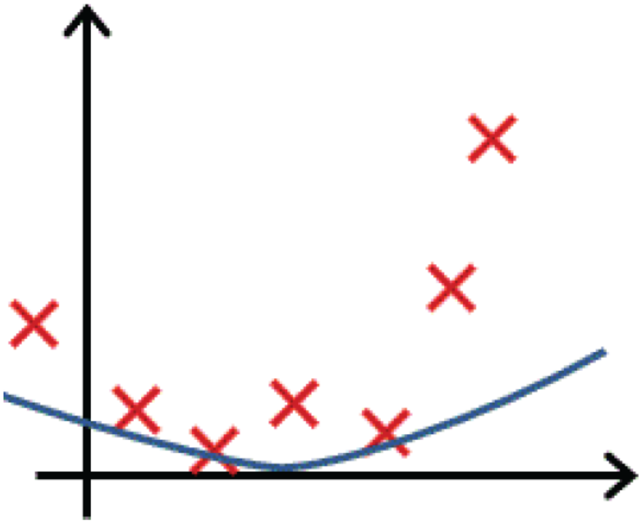
a)



b)



c)



d)

Solution: B