Supervision Assignments in MLRW - Set 1

Dionysis Manousakas

February 23, 2017

1. (*Binary classifier*)

   i. Assuming a binary classification problem, which is the maximum error that any possible dataset can have? Give an argument for your answer.

   ii. Think of an example of a binary classification problem where it seems difficult to train a classifier able to further reduce this error beyond the theoretical bound found in i.

2. (*Overfitting / Model evaluation*)

   i. Is a classifier trained on less training data less likely to overfit?

   ii. Given m data points, does the training error converge to the true error as $m \to \infty$?

   iii.You are a reviewer for an international Machine Learning conference. Would you accept or reject each paper? Provide a one sentence justification.

   - My algorithm is better than yours. Look at the training error rates!

   - My algorithm is better than yours. Look at the test error rates! (Footnote: reported results for parameter $\lambda = 0.4724332783241324108 3478.$)

   - My algorithm is better than yours. Look at the test error rates! (Footnote: reported results for best value of λ .)

   - My algorithm is better than yours. Look at the test error rates! (Footnote: reported results for best value of λ , chosen with 10-fold cross validation.)

3. (*Bayes rule for medical diagnosis*)

   After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? Show your calculations as well as giving the final result. (Murphy's book Exercise 2.4)

4. (*Bayes rule)*)

   i. Consider a classification problem with two classes and n binary attributes. How many parameters would you need to learn with a Naive Bayes classifier?

   ii. Suppose we have a set of k hypotheses (or models)

   $$h = [h_1, h_2, ..., h_k]$$

   for an observed dataset $\mathcal{D}$. According to the Bayes rule

   $$p(h_i|\mathcal{D}) \propto p(\mathcal{D}|h_i)p(h_i)$$

   for $i \in 1, 2, ...k$.

   Write the more flexible and least flexible prior distributions we can assume over the given hypotheses' set. How can we make the least flexible prior less rigid?

5. (*The Monty Hall problem*)

   On a game show, a contestant is told the rules as follows: There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other two doors, and he will do so in such a way as not to reveal the prize. For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door. Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially, the prize is equally likely to be behind any of the 3 doors. Hint: use Bayes rule. (Murphy's book Exercise 2.5)

6. (*Zipf's law and other real-data distributions*)

   i. Describe a probabilistic process that generates data which follow the normal distribution.

   ii. Describe a probabilistic process that generates data which follow the Zipf's law (or in other terms a power law distribution).

   iii. Describe a real-world data set which you believe could be modelled using a mixture of distributions. Argue why a mixture model is a sensible model for your real world data set. What do you expect the mixture components to represent? How many components (or clusters) do you think there would be? What parametric form would each component have?

7. (*Cross-validation*)

   i. In n-fold cross-validation each data point belongs to exactly one test fold, so the test folds are independent. Are the error estimates of the separate folds also independent? So, given that the data in test folds i and j are independent, are e_i and e_j, the error estimates on test folds i and j , also independent? Is there an a priori good choice of n for n-fold cross-validation?

   ii. Assume that we are using some classifier of fixed complexity. Draw a graph showing two curves: test error vs. the number of training examples and cross-validation error vs. the number of training examples.

   iii. Assume that we are using classifiers of increasing complexity. Draw a graph showing three curves: test error vs. complexity, cross-validation error vs. complexity and training error vs. complexity.

8. (*Discriminative vs Generative classifiers*)

   i. Your billionaire friend needs your help. She needs to classify job applications into good/bad categories, and also to detect job applicants who lie in their applications using density estimation to detect outliers. To meet these needs, do you recommend using a discriminative or generative classifier? Why?

   ii. Your billionaire friend also wants to classify software applications to detect bug-prone applications using features of the source code. This pilot project only has a few applications to be used as training data, though. To create the most accurate classifier, do you recommend using a discriminative or generative classifier? Why?

   iii. Finally, your billionaire friend also wants to classify companies to decide which one to acquire. This project has lots of training data based on several decades of research. To create the most accurate classifier, do you recommend using a discriminative or generative classifier? Why?

9. (*Pointwise Mutual Information*)

   Pointwise mutual information (PMI) is a measure of statistical independence which can tell use whether two words (or more generally, two statistical events) tend to occur together or not.The PMI between two events x and y is defined as

$$PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

   Lets consider two examples:

   - x is "eat is the first word in a bigram" and y is "pizza is the second word in a bigram".
   - x is "happy occurs in a Tweet" and y is "pizza occurs in a Tweet".

i. For each example, what does $P(x, y)$ represent?

ii. What do negative, zero, and positive PMI values represent in terms of the statistical independence of x and y? (Hint: consider what must be true of the relationship between $P(x, y)$ and $P(x)P(y)$ for the PMI to be negative, zero, or positive.) Give some example pairs of words that you would expect to have negative or positive PMI (in either the bigram or Tweet scenario).