Supervision Assignments in MLRW - Set 1

Dionysis Manousakas

March 18, 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.

   answer

   Let's randomly map our data to the two classes. This will give us accuracy α. If $\alpha < 0.5$ at zero cost we can obtain acccuracy $1 - \alpha \geq 0.5$ by reverting the assigmnet to the two classes. This assumes that we have access to a black box that give us back the accuracy of the classification which might be unclear from the description of the question...

   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.

   answer

   Fair dice.

2. (*Overfitting / Model evaluation*)

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

   answer

   More likely to overfit, since less data increase the probability that the classifier will start learning noise

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

   answer

   Yes, if they are i.i.d.

   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!

   answer

   Reject - the training error is optimistically biased.

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

   answer

   Reject - A λ with 15 decimal places suggests a highly tuned solution, probably looking at the test data.

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

   anser

   Reject - Choosing λ based on the test data?

- 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.)

answer

Accept - Cross validation is the appropriate method for selecting parameters.

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)

answer

$$\begin{aligned} P(disease|test) &= P(test|disease)P(disease)/P(test) \\ &= \frac{P(test|disease)P(disease)}{P(test|disease)P(disease) + P(test|notdisease)P(notdisease)} \\ &= 0.99 * 0.0001/(0.99 * 0.0001 + 0.01 * 0.9999) \approx 0.009804 \end{aligned} \tag{1}$$

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?

answer:

NB has $1 + 2n$ parameters :

- prior $P(y = T)$, and
- for every attribute x_i,we have $p(x_i = T|y_i = T)$ and $p(x_i = T|y_i = F)$.

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?

answer

The most flexible is assigning a uniform prior over the classes (we do not favor a priori any of the hypotheses and let the data speak for them), while the most rigid is using a delta (or dirac) prior, which assigns the entire weight of the prior to only one class (and zero to the rest of the classes, which implies zero posterior regarding of the likelihoods). This can become less rigid by smoothing, which will result to giving a minimum prior to each one of the classes.

`Cromwell's Rule`

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)

answer

Let's solve it using Bayes' rule, although there are simpler ways to approach it...

For simplicity let:

- D3: The Event of Monty Hall opening door 3.
- C1: The Event of finding the prize behind door 1.
- C2: The Event of finding the prize behind door 2.
- C3: The Event of finding the prize behind door 3.

The prior probability of Monty Hall finding the prize behind any door is obviously $P(C1) = P(C2) = P(C3) = 1/3$.

Assume you chose the door No.1. So let's focus here on the 3rd Door!

The probability that Monty Hall opens door No.3. given the car is behind door No.1 is: $p(D3|C1) = 1/2$.

We know also that Monty Hall will never open the door which has the car so the probability that Monty Hall opens door No.3. given that the car is behind door No.3. is: $p(D3|C3) = 0$.

Contrary, the probability that Monty Hall opens door No.3. given the car is behind door No.2. is: $p(D3|C2) = 1$.

Now to the Bayes's part:

$$p(C1|D3) = p(D3|C1) * p(C1)/p(D3) = (1/2 * 1/3)/(1/2) = 1/3$$

$$p(C2|D3) = p(D3|C2) * p(C2)/p(D3) = (1 * 1/3)/(1/2) = 2/3$$

therefore, if you choose No.1. initially and Monty Hall opens door No.3. to reveal that it has no car, the probability of a car standing behind door No.2. is 2/3. Therefore you should *always switch* your selection of your initial door.

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

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

answer

The key to understanding the applications of the normal distribution is understanding the Central Limit Theorem. The key idea is that if the random quantity you are trying to understand is the result

of the accumulation of a large number of (approximately) independent random variations that are (approximately) the same size, then that quantity will (approximately) follow a normal distribution.

Suppose you commute to work taking the bus every morning. Barring a serious delay due to an accident or breakdown, there's a good chance that the time your bus arrives to pick you up in the morning will be well-modelled by a normal distribution. Why? Because the arrival time will depend on things like times the bus spends at traffic lights and times it spends waiting for other students before it gets to you, and these individual variations are probably roughly independent and roughly similar in duration. Now if you write down the arrival time of the bus every day for a couple of months (or really even less), you can accurately estimate the mean and standard deviation of the arrival time.

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

answer

A discrete-time stochastic process generated heavy tailed data is the Chinese Restaurant Process. Imagine a Chinese restaurant with an infinite number of circular tables, each with infinite capacity. Customer 1 sits at the first table. The next customer either sits at the same table as customer 1, or the next table. This continues, with each customer choosing to sit at an occupied table with a probability proportional to the number of customers already there (i.e., they are more likely to sit at a table with many customers than few), or an unoccupied table. At time n, the n customers have been partitioned among $m \leq n$ tables (or blocks of the partition). The results of this process are exchangeable, meaning the order in which the customers sit does not affect the probability of the final distribution.

In the context of network science, networks created using the preferential attachment (or Barabasi-Albert) model display degree distributions which follow the Zipf's law. Preferential attachment means that the more connected a node is, the more likely it is to receive new links. Nodes with higher degree have stronger ability to grab links added to the network.

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?

answer

Imagine that we are given an $N \times N$ black-and-white image that is known to be a scan of a hand-written digit between 0 and 9, but we don't know which digit is written. We can create a mixture model with $K = 10$ different components, where each component is a vector of size N^2 of Bernoulli distributions (one per pixel).

7. (*Cross-validation*)

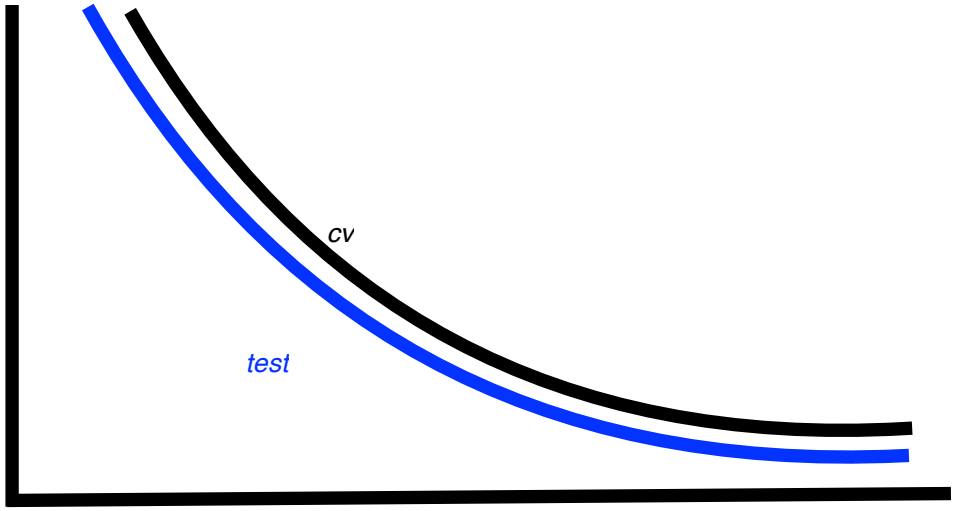
i. In n-fold cross-validation each data point belongs to exactly one test fold, so the test folds are independent. Are the error estimates of the separate folds also independent? So, given that the data in test folds i and j are independent, are e_i and e_j, the error estimates on test folds i and j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

answer

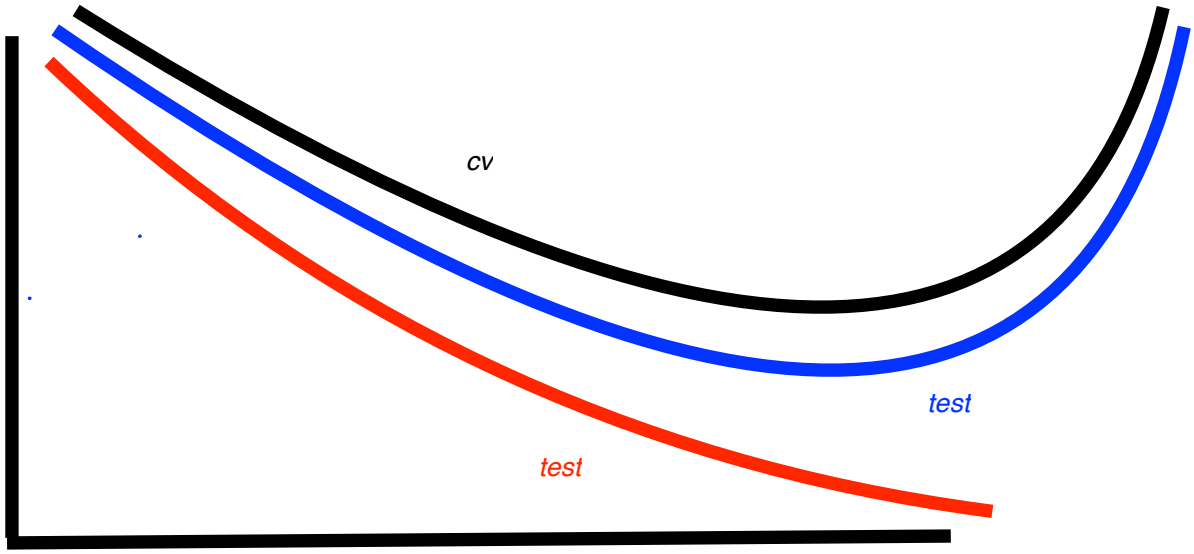
False. Since a data point appears in multiple folds the training sets are dependent and thus test fold error estimates are dependent.

There is no a priori good choice for the number of folds in cross-validation. We do not know the relation between sample size and the accuracy. High n increases correlation in training set and decreases variance of estimates. How much depends on the data and the learning method. However, in any case high n allows a larger training set which is desirable.

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?

answer Generative, since for density estimation we need $p(x|y)$.

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?

answer Discriminative. We have only few datapoints, thus we'd rather not do any further assumptions on how the data are generated and use full expressive power of our model to distinguish the bug-prone applications.

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?

answer Generative. We have lots of training data, thus we can impose reasonable assumptions on how the data are generated and train a generative classifier. Placing some structure on the data will also help us prevent potential issues of overfitting.

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?

answer:

The probability of the bigram eat pizza, or the probability of using the words happy and pizza in the same Tweet.

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).

answer:

- Negative: x and y are less likely to occur together than if independent. Ex: treaty and pizza, two words which are very unlikely to be used together in the same conversation/text, much less in a single Tweet or bigram.
- Zero: x and y are independent.
- Positive: x and y are more likely to occur together than if independent. Ex: pepperoni and pizza.