Supervision Assignments in MLBI
Lent Term 2018
Set 1

Supervisor : Dionysis Manousakas
dm754@cam.ac.uk

February 10, 2018

1. (*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.)

2. (*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)

3. (*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 uninformed and most informed prior distributions we can assume over the given hypotheses' set. How can we make the most informed prior less rigid?

4. (*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)

5. (*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.

6. (*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?

7. (*Models for binary vectors*)

   Consider a data set of binary (black and white) images. Each image is arranged into a vector of pixels by concatenating the columns of pixels in the image. The data set has N images $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$ and each image has D pixels, where D is (number of rows $\times$ number of columns) in the image. For example, image $\mathbf{x}^{(n)}$ is a vector $(x_1^{(n)}, ..., x_D^{(n)})$ where $x_d^{(n)} \in \{0, 1\}$ for all $n \in \{1, ..., N\}$ and $d \in \{1, ..., D\}$.

   (a) Explain why a multivariate Gaussian would not be an appropriate model for this data set of images.

   (b) Assume that the images were modelled as independently and identically distributed samples from a D-dimensional **multivariate Bernoulli distribution** with parameter vector $\mathbf{p} = (p_1, ..., p_D)$, which has the form

$$P(\mathbf{x}|\mathbf{p}) = \prod_{d=1}^{D} p_d^{x_d}(1 - p_d)^{(1-x_d)}$$

   where both $\mathbf{x}$ and $\mathbf{p}$ are D-dimensional vectors.

   i. What is the equation for the maximum likelihood (ML) estimate of $\mathbf{p}$? Note that you can solve for $\mathbf{p}$ directly.

   ii. Assuming independent Beta priors on the parameters p_d

$$P(p_d) = \frac{1}{B(\alpha, \beta)} p_d^{\alpha-1}(1 - p_d)^{\beta-1}$$

   and $P(\mathbf{p}) = \prod_d P(p_d)$, What is the maximum a posteriori (MAP) estimate of $\mathbf{p}$? Hint: maximise the log posterior with respect to $\mathbf{p}$.

(c) Download the dataset `binarydigits.txt`, which contains $N = 100$ images with $D = 64$ pixels each, in an $N \times D$ matrix. These pixels can be displayed as 8×8 images by rearranging them.. View them in Matlab using the script `bindigit.m` (almost no Matlab knowledge required to do this).

i. Write code to learn the ML parameters of a multivariate Bernoulli from this data set and display these parameters as an 8×8 image.

ii. Modify your code to learn MAP parameters with $\alpha = \beta = 3$. What is the new learned parameter vector for this data set? Explain why this might be better or worse than the ML estimate.

8. (*Model selection*)

   In the binary data model above, write down the expressions needed to calculate the (relative) probability of the three different models:

   (a) all D components are generated from a Bernoulli distribution with $p_d = 0.5$

   (b) all D components are generated from Bernoulli distributions with unknown, but identical, p_d

   (c) each component is Bernoulli distributed with separate, unknown p_d

   Assume the prior probability of all three models is the same. Which model is the most likely given the data in `binarydigits.txt`?

9. (*Dependence of p-values on irrelevant information -or why we should become Bayesians :-) (from McKay's book)*

   (a) In an expensive laboratory, Dr. Bloggs tosses a coin labelled a and b twelve times and the outcome is the string

   $$aaabaaaabaab,$$

   which contains three bs and nine as. What evidence do these data give that the coin is biased in favour of a? Is it significant at the level of 5% ?

   (b) Dr. Bloggs pays careful attention to the calculation of 1.1, and responds *'no, no, the random variable in the experiment was not the number of bs: I decided before running the experiment that I would keep tossing the coin until I saw three bs; the random variable is thus the total number of tosses, n'.* A different calculation is required in order to assess the 'significance' of the result $n = 12$. Now, the probability distribution of n given $\mathcal{H}_0$ is the probability that the first $n-1$ tosses contain exactly $r-1$ bs and then the nth toss is a b. What evidence do these data give that the coin is biased in favour of a in this case? Is the evidence significant at the level of 5% ?