DEGREE OF MASTER OF SCIENCE

MSc in Computer Science

Machine Learning

Hilary Term 2016

Your answer to this paper is to be handed in at the Examination Schools, High Street, by 12 noon on 18th April 2016

The assignment should be submitted in an envelope, clearly marked with your candidate number, (but not your name) and the name of the course, and addressed to the Chairman of Examiners, MSc in Computer Science.

NB: You must NOT discuss this examination paper with anyone.

This paper contains 3 questions; candidates should attempt all questions. There is a total of 100 marks available for this paper.

Instructions to candidates: You may make use of books, class slides, and online resources for revision. However, you are not allowed to discuss with anybody and you are not allowed to search online directly for answers to the questions in this examination.

Question 1

Support Vector Machines and False Positive Errors

You work for an email service provider and you've recently received several complaints from users that genuine emails are being directed to their spam folders. For this problem you should assume that useful features have already been extracted to map each email as a vector $\mathbf{x} \in \mathbb{R}^n$. You are using support vector machines (with linear kernels) to classify emails as spam or not. Thus, your model consists of parameters $(\mathbf{w}, w_0)$ and on input vector $\mathbf{x} \in \mathbb{R}^n$, the output is given by:

$$\hat{y} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} + w_0 \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

The training data you are using is $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^m$, where $y_i = 1$ indicates that the email is **spam**. For a new input $\mathbf{x}_{\text{new}}$, if your model predicts $\hat{y}_{\text{new}} = 1$ it will go to the user's spam folder. For an input $\mathbf{x}$ with true label $y = -1$ (true label **not spam**), we will say that the model made a *false positive error* if the prediction was $\hat{y} = 1$ (predicted **spam**). At this point you are greatly concerned with reducing the false positive errors of your model.

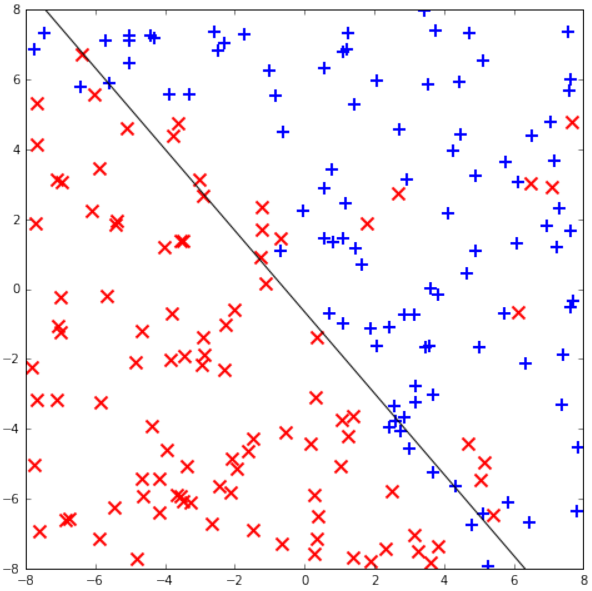


Figure 1: The '+' denote positive examples ($y = 1$) and 'x' denote negative examples ($y = -1$).

(a) One of your colleagues comes along and suggests a simple modification that requires no re-training of the classifier. Simply increase the value of w_0 to be $w_0 + \alpha$, for $\alpha > 0$. Now, you will classify fewer emails as spam. Even better, you could quickly choose the best value of α by cross-validation: choose the value that gives the smallest false positive error rate. Give up to two reasons for not agreeing with your colleague. (10 marks)

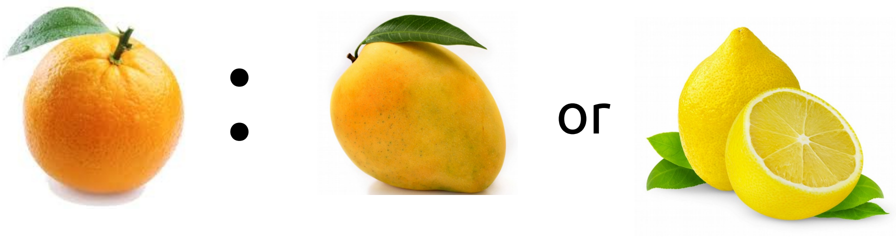
(b) Another colleague suggests a different line of attack: "Make 200 copies of every *spam* email, so that the training data has a lot more spam emails than it had earlier. Then just retrain the SVM.". Describe if you agree with your colleague—and make any changes to their solution if you think necessary. Also explain how the extra copies need not slow down the training time for your SVM classifier. (15 marks)

(c) Finally, your boss tells you: “We really really need to avoid genuine emails going to the spam folder. If that means a lot of spam makes its way to the inbox, so be it. Explain how you would re-formulate the SVM objective and constraints so that there will be *no* false positive errors on the training set. Of course, you still want to minimise the false negative errors as much as you can. (10 marks)

Question 2

Clustering Fruits

You want to cluster fruit. One option would be to go to your plant biologist friends down the road and ask them to come up with detailed features describing every fruit. But, it’s a nice summer day and you think you’ve got a better idea: you’ll simply go to the market and ask random people to rate how similar two fruits are on a scale of 1 to 10. You begin with an orange and a grapefruit and soon realise the problem—you’ve received scores between 4-9 for the same pair of fruits.



Luckily, you brought some friends along who suggest a way to get more consistent responses: take three fruits, say an orange (a), then an apple (b) and a grapefruit (c), then ask the question is (a) more similar to (b) or (c), *i.e.,* is an orange more like an apple or a grapefruit? Though even now, there could be some inconsistencies. For example if a = orange, b = mango and c = lemon, one might say an orange is more like a mango due to its colour, but equally one could argue that it’s more like a lemon because they are both citrus fruits!

You have a total of n fruits and you also enforce an arbitrary order on them, say alphabetic; we will use $a \prec b$ to denote that a appears before b according to this order. You design a model that for every triple $(a : b, c)$ of (distinct) fruits, assigns a probability $p^a_{b,c}$ to indicate the probability that a (random) person would say a is more similar to b than c. Note that you must have $p^a_{b,c} + p^a_{c,b} = 1$, which limits the degrees of freedom in the model. The input data (by design) consists of triples $(a : b, c)$, where $b \prec c$.[1] The output is 1 if the person thought a was more similar to b than c and 0 otherwise. Thus, your training data is of the form $\langle ((a_i : b_i, c_i), y_i) \rangle_{i=1}^m$, where $b_i \prec c_i$ for all i and $y_i \in \{0, 1\}$.

(a) For the model described above, we will use the parameter $p^a_{b,c}$ when $b \prec c$ to denote the probability that a (random) person considers a more like b than c, and $(1 - p^a_{b,c})$ to denote

[1] This may not be good practice for surveys since people may be biased by the positioning.

the probability that the person may think that a is more similar to c than b. Write the likelihood of observing the data $\langle((a_i : b_i, c_i), y_i)\rangle_{i=1}^m$ described above given the parameters $p^a_{b,c}$. You may assume that the y_i are independent for all the datapoints, since you are genuinely picking random people in the street. How many (free) parameters does your model have? (10 marks)

(b) Unfortunately this approach involves a lot of parameters. Moreover, it's hard to know how to go from these probabilities to actual similarity measures between fruits to use for clustering. An alternative is to use $M_{a,b}$ to denote similarity between a and b. We define these parameters $M_{a,b}$ for all $a \prec b$ (as $M_{b,a} = M_{a,b}$). Furthermore, we set $M_{a,a} = 1$ for all a, and require that $M_{a,b} \leq 1$ for all $a \neq b$. We have reduced the number of parameters to $\binom{n}{2}$. We model the probability that a (random) user thinks a is more like b than c as $p^a_{bc} = \frac{\exp(M_{a,b})}{\exp(M_{a,b})+\exp(M_{a,c})}$. Write the negative log likelihood in terms of the parameters $M_{a,b}$ and explain how you would solve the resulting optimisation problem. (15 marks)

(c) Actually, it turns out that the above optimisation problem can be made to return $M_{a,b}$, such that the resulting $n \times n$ matrix is positive semi-definite. Explain how you would use the matrix M to cluster the fruits. (10 marks)

Question 3

Deep Residual Networks and Highway Networks

For this part you should read the following two articles:

(1) Deep Residual Learning for Image Recognition. Available at `http://arxiv.org/abs/1512.03385`

(2) Training Very Deep Networks. Available at `http://arxiv.org/abs/1507.06228`

Based on your reading answer the following questions.

(a) What do the authors of (1) mean by the degradation problem? If you perform linear regression with polynomial basis expansion (adding the terms of $x^2, x^3, \ldots$) would you expect to have a similar problem as you increase the degree? Explain why this may be occurring in one or both of the cases. (10 marks)

(b) Explain in what ways the models proposed in the two papers are similar. Could you view one approach as being a special case of the other? (10 marks)

(c) Explain what the authors of (2) mean by *lesioning*. What aspects of their model do the authors explain using *lesioning*? (10 marks)

(d) *(Not for credit)* You may use this space to add any additional comments about the two papers. While, there is no credit assigned for the question, this will help you formulate your thoughts about these papers. (0 marks)