COMP9417 - Machine Learning
Tutorial: Classification

Weekly Problem Set: Please submit questions 2c, 3a and 3b on Moodle by 12pm Tuesday 8th March, 2022. Please only submit these requested questions and no others.

Question 1 (Bayes Rule)

Assume that the probability of a certain disease is 0.01. The probability of testing positive given that a person is infected with the disease is 0.95, and the probability of testing positive given that the person is not infected with the disease is 0.05.

(a) Calculate the probability of testing positive.

(b) Calculate the probability of being infected with the disease, given that the test is positive.

(c) Now assume that you test the individual a second time, and the test comes back positive (so two tests, two positives). Assume that conditional on having the disease, the outcomes of the two tests are independent, what is the probability that the individual has the disease? (note, conditional independence in this case means that $P(TT|D) = P(T|D)P(T|D)$, and not $P(TT) = P(T)P(T)$.)

Question 2 (Lecture Review)

In this question, we will review some important ideas from the lecture.

(a) What is probabalistic classification? How does it differ from non-probabalistic classification methods?

(b) What is the Naive Bayes assumption and why do we need it?

(c) Consider the problem from lectures of classifying emails as spam or ham, with training data summarised below: Each row represents an email, and each email is a combination of words taken

e_1	b	d	e	b	b	d	e			
e_2	b	c	e	b	b	d	d	e	c	c
e_3	a	d	a	d	e	a	e	e		
e_4	b	a	d	b	e	d	a	b		
e_5	a	b	a	b	a	b	a	e	d	
e_6	a	c	a	c	a	c	a	e	d	
e_7	e	a	e	d	a	e	a			
e_8	d	e	d	e	d					

from the set $\{a, b, c, d, e\}$. We treat the words d, e as stop words - these are words that are not useful for classification purposes, for example, the word 'the' is too common to be useful for classifying documents as spam or ham. We therefore define our vocabulary as $V = \{a, b, c\}$. Note that in this case we have two classes, so $k = 2$, and we will assume a uniform prior, that is:

$$p(c_+) = p(c_-) = \frac{1}{2},$$

where $c_+ =$ spam, $c_- =$ ham. Review the multivariate Bernoulli Naive Bayes set-up and classify the test example: assume we get a new email that we want to classify: $e_\star = abbdebb$

(d) Next, review Smoothing for the multivariate Bernoulli case. Why do we need smoothing What happens to our previous classification under the smoothed multvariate Bernoulli model?

(e) Redo the previous analysis for the Multinomial Naive Bayes model without smoothing. Use the following test email: $e_\star = abbdebbcc$

(f) Repeat the analysis for the smoothed Multinomial Naive Bayes model.

Question 3. Binary Logistic Regression, two perspectives

Recall from previous weeks that we can view least squares regression as a purely optimisation based problem (minimising MSE), or as a statistical problem (using MLE). We now discuss two perspectives of the Binary Logistic Regression problem. In this problem, we are given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where the x_i's represent the feature vectors, just as in linear regression, but the y_i's are now binary. The goal is to model our output as a probability that a particular data point belongs to one of two classes. We will denote this predicted probability by

$$P(y = 1|x) = p(x)$$

and we model it as

$$\hat{p}(x) = \sigma(\hat{w}^T x), \qquad \sigma(z) = \frac{1}{1 + e^{-z}},$$

where $\hat{w}$ is our estimated weight vector. We can then construct a classifier by assigning the class that has the largest probability, i.e.:

$$\hat{y} = \arg\max_{k=0,1} P(\hat{y} = k|x) = \begin{cases} 1 & \text{if} \quad \sigma(\hat{w}^T x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

note: do not confuse the function $\sigma(z)$ with the parameter σ which typically denotes the standard deviation.

(a) What is the role of the logistic sigmoid function $\sigma()$ in the logistic regression formulation? Why are we not able to simply use linear regression here? (a plot of $\sigma(z)$ may be helpful here).

(b) We first consider the statistical view of logistic regression. Recall in the statistical view of linear regression, we assumed that $y|x \sim N(x^T\beta^*, \sigma^2)$. Here, we are working with binary valued random variables and so we assume that

$$y|x \sim \text{Bernoulli}(p^*), \qquad p^* = \sigma(x^T w^*)$$

where $p^* = \sigma(x^T w^*)$ is the true unknown probability of a response belonging to class 1, and we assume this is controlled by some true weight vector w^*. Write down the log-likelihood of the data D (as a function of w), and further, write down the MLE objective (but do not try to solve it).

(c) An alternative approach to the logistic regression problem is to view it purely from the optimisation perspective. This requires us to pick a loss function and solve for the corresponding minimizer. Write down the MSE objective for logistic regression and discuss whether you think this loss is appropriate.

(d) (**optional**) Consider the following problem: you are given two discrete probability distributions, P and Q, and you are asked to quantify how far Q is from P. This is a very common task in statistics and information theory. The most common way to measure the discrepancy between the two is to compute the Kullback-Liebler (KL) divergence, also known as the relative entropy, which is defined by:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)},$$

where we are summing over all of the possible values of the underlying random variable. A good way to think of this is that we have a true distribution P, an estimate Q, and we are trying to figure out how bad our estimate is. Write down the KL divergence between two bernoulli distributions $P = \text{Bernoulli}(p)$ and $Q = \text{Bernoulli}(q)$.

(e) (**optional**) Continuing with the optimisation based view: In our set-up, one way to quantify the discrepancy between our prediction $\hat{p}_i$ and the true label y_i is to look at the KL divergence between the two bernoulli distributions $P_i = \text{Bernoulli}(y_i)$ and $Q_i = \text{Bernoulli}(\hat{p}_i)$. Use this to write down an appropriate minimization for the logistic regression problem.

(f) (**optional**) In logistic regression (and other binary classification problems), we commonly use the cross-entropy loss, defined by

$$\mathcal{L}_{XE}(a,b) = -a \ln \frac{a}{b} - (1-a) \ln \frac{1-a}{1-b}.$$

Using your result from the previous part, discuss why the XE loss is a good choice, and draw a connection between the statistical and optimisation views of logistic regression.