MLPR Assignment

Alexander McMurray
Student Number: 1367329

November 13, 2013

1 Question 1 *(10 marks)*

(a) Describe mathematically the following two probabilistic models:
(i) A beta-Bernoulli model that assumes that the probability of a tweet being retweeted is the same for every day of the week.

Let y represent whether or not the tweet was retweeted (i.e. $y = 1$ for retweeted, $y = 0$ for not retweeted) and x_i represent whether it was tweeted on a particular day (i.e. $x_1 =$ Sunday, $x_2 =$ Monday...) then assuming that the probability of retweeting is the same for every day of the week we can state that the probability of retweeting is some value, θ. i.e.:

$$P(y = 1) = \theta \text{ and therefore } P(y = 0) = 1 - \theta \tag{1.1}$$

where θ is drawn from the Beta distribution i.e. $\theta \sim \text{Beta}(\alpha, \beta)$.

Therefore we can state our prior on θ as:

$$P(\theta) = \frac{1}{\text{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{1.2}$$

where

$$\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{1.3}$$

is the Beta function and the Gamma function is given by:

$$\Gamma(n) = (n-1)!, \; n \in \mathbb{N} \tag{1.4}$$

i.e. n is a natural number, $n = 1, 2, 3, 4 \ldots$

In practice however it is better to use MATLAB's `beta` or `betaln` functions to avoid numerical errors.

As we don't know much about θ it would be prudent to set $\alpha = \beta = 2$. (The reason for using two as the value rather than one (the uniform prior) will be explained in part (e).)

The likelihood is given by a Bernoulli distribution:

$$P(X|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{1.5}$$

where n is the total number of tweets over all days (as we are assuming all days are equally likely so the day doesn't matter) and k is the number of retweeted tweets over all day.

Then we can calculate the posterior via Bayes' theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{1.6}$$

where $P(X)$ can be obtained by marginalising out θ from $P(X|\theta)$ i.e.:

$$P(X) = \int_0^1 P(X|\theta)P(\theta)d\theta = \binom{n}{k}\frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 \theta^{k+(\alpha-1)}(1-\theta)^{(n-k)+(\beta-1)}d\theta \tag{1.7}$$

Note that we have essentially added *pseudocounts* of $\alpha - 1$ to the number of positive cases observed and $\beta - 1$ to the number of negative counts observed.

The solution to such an integral is given by the Beta function:

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \quad \text{for } \mathrm{Re}(x),\ \mathrm{Re}(y) > 0 \tag{1.8}$$

Applying Equation (1.8) to Equation (1.7) and using the properties given in Equations (1.3) and (1.4) we obtain that:

$$P(X) = \binom{n}{k}\frac{\mathrm{B}(k+\alpha,(n-k)+\beta)}{\mathrm{B}(\alpha,\beta)} \tag{1.9}$$

This can be evaluated as we know n and k from the data and we have chosen values for α and β.

So via Bayes Theorem (Equation (1.6)) we obtain the posterior:

$$P(\theta|X) = \frac{\theta^{k+(\alpha-1)}(1-\theta)^{(n-k)+(\beta-1)}}{\mathrm{B}(k+\alpha,(n-k)+\beta)} \tag{1.10}$$

Which can be plotted (as a function of θ) to investigate which values of θ (the probability of a tweet being retweeted) are most likely. The maximum of the distribution gives the *Maximum a Posteriori* (MAP) estimate.

(ii) A model that uses separate beta-Bernoulli distributions for each day of the week and assumes the probability of a tweet being retweeted is different for every day of the week.

If we now assume that θ may be different for each day then we can use separate beta-Bernoulli distributions for each day of the week where:

$$P(y_i = 1) = \theta_i \text{ and therefore } P(y_i = 0) = 1 - \theta_i \tag{1.11}$$

where $i \in \{1, 2, 3 \dots 7\}$ corresponding to each day (i.e. 1=Sunday, 2=Monday etc.) so y_i is whether it a tweet was retweeted on a certain day and θ_i is the probability of it being retweeted on a particular day.

Then we can use the same approach described in part (a)(i) on the distribution for each day to obtain the posterior distribution:

$$P(\theta_i|X_i) = \frac{\theta_i^{k_i+(\alpha-1)}(1-\theta_i)^{(n_i-k_i)+(\beta-1)}}{\mathrm{B}(k_i+\alpha,(n_i-k_i)+\beta)} \tag{1.12}$$

where X_i is the data for a given day, n_i is the total number of tweets on a given day and k_i is the number of retweeted tweets on a given day.

(b) Using the second model extract the tweets that were sent on Monday. Plot on a single figure the posterior distribution for Monday after incorporating the first 1, 100 and 1000 data points. Use a Beta prior with values $\alpha = 2$, $\beta = 2$.
(i) How does the posterior change as more data points are added?
(ii) Why does this happen?

This means we want to plot:

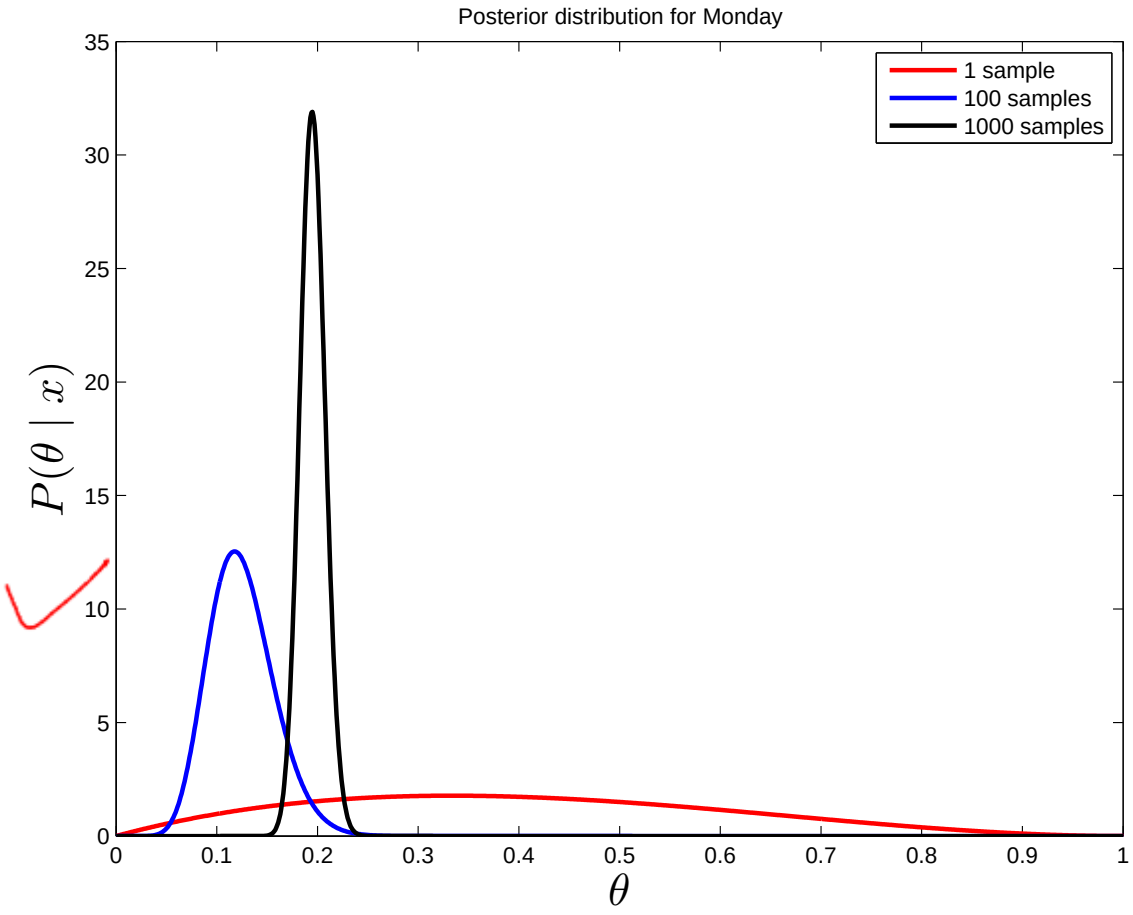
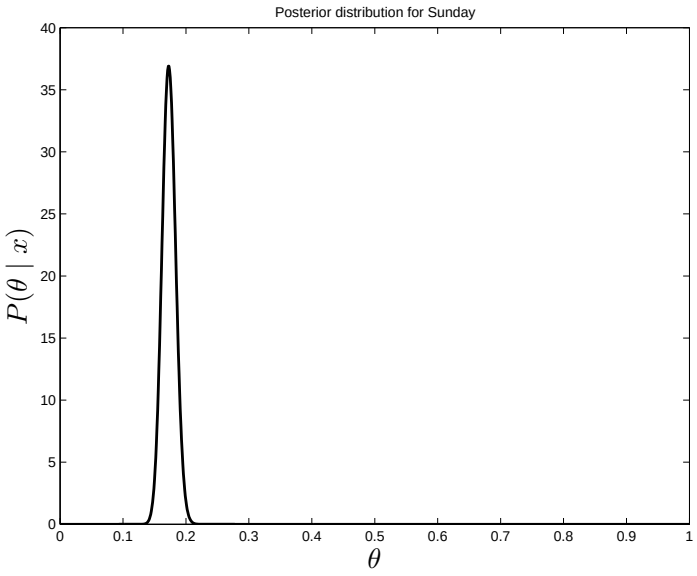


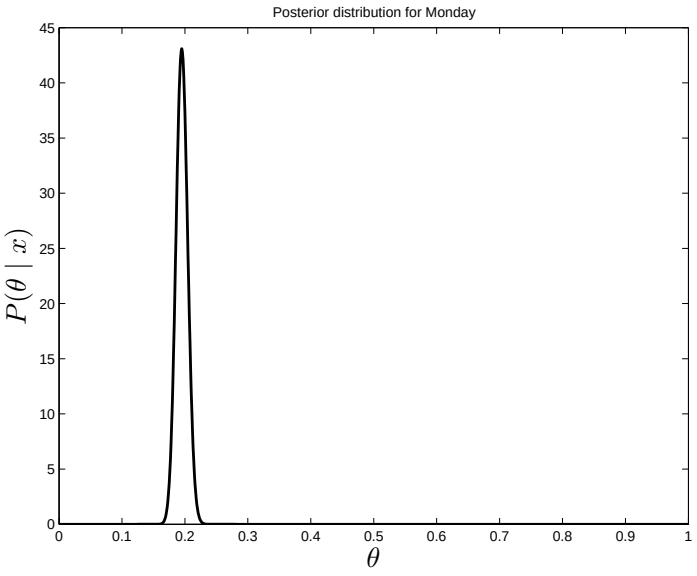
Figure 1.1: The posterior distribution for Monday incorporating the first 1, 100 and 1000 data points. $P(\theta; \alpha, \beta) = \text{Beta}(\theta; 2, 2)$ was used as the prior.

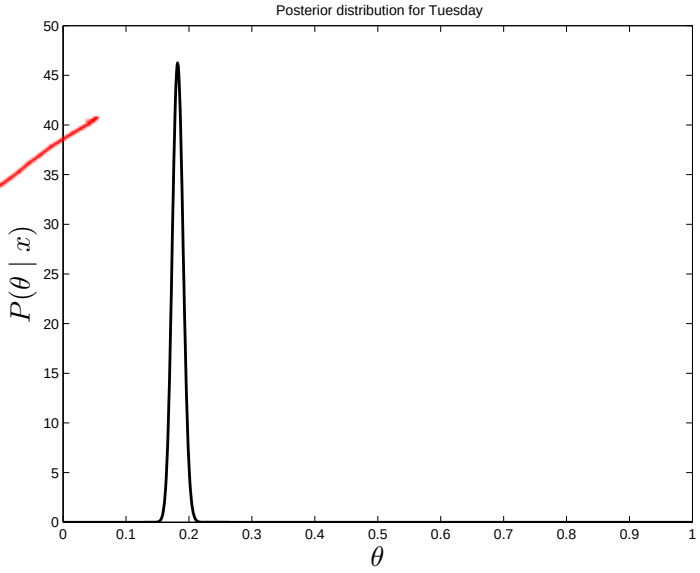
$$P(\theta_2|X_2) = \frac{\theta_2^{k_2+1}(1-\theta_2)^{(n_2-k_2)+1}}{\text{B}(k_2+2, (n_2-k_2)+2)} \tag{1.13}$$

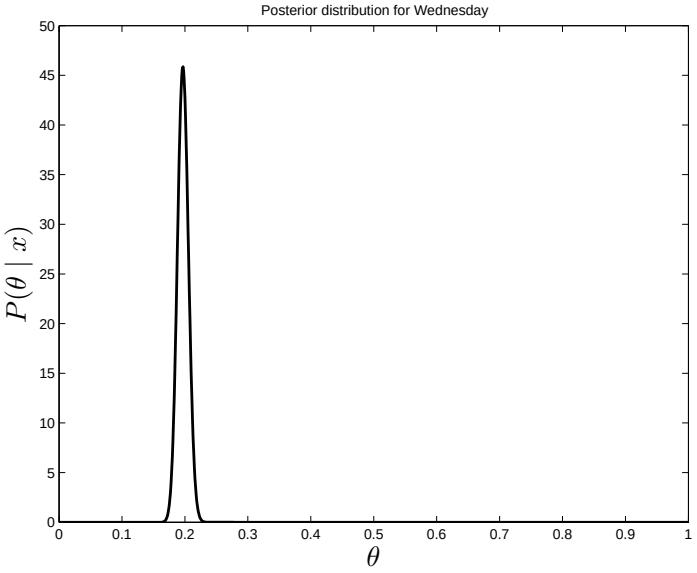
As more data points are added the posterior becomes narrower and higher and shifts position. This happens because the more data we have, the more information we have and thus we are less reliant on our assumptions (represented by the prior). With more data the distribution approaches a delta function located at the *Maximum Likelihood Estimate* (MLE) because the prior becomes negligible compared to the data (i.e. the data swamps the prior). For large enough data samples the MAP estimate $\approx$ MLE and the posterior is independent of the prior as stated by Bernstein-von Mises Theorem.

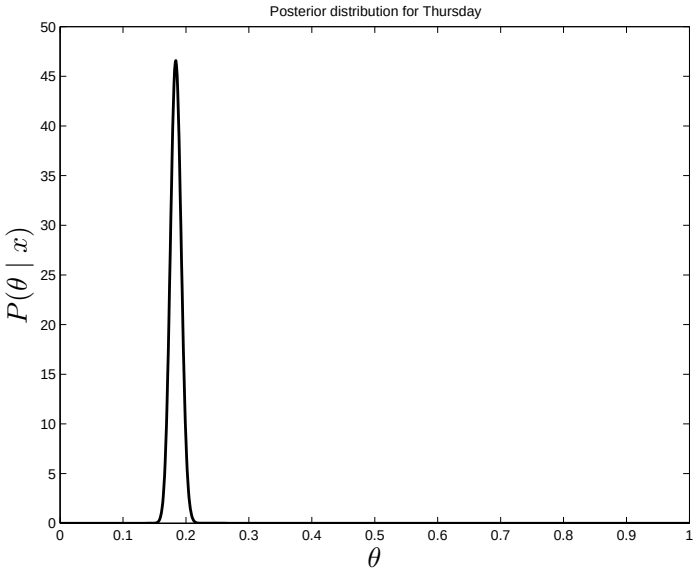
(c) Plot the posterior distribution after all data points have been incorporated for each of the seven days of the week (on separate figures).

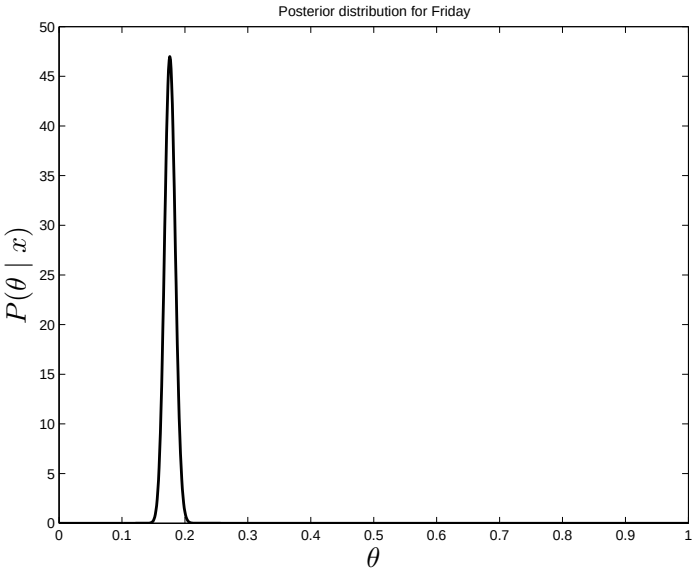


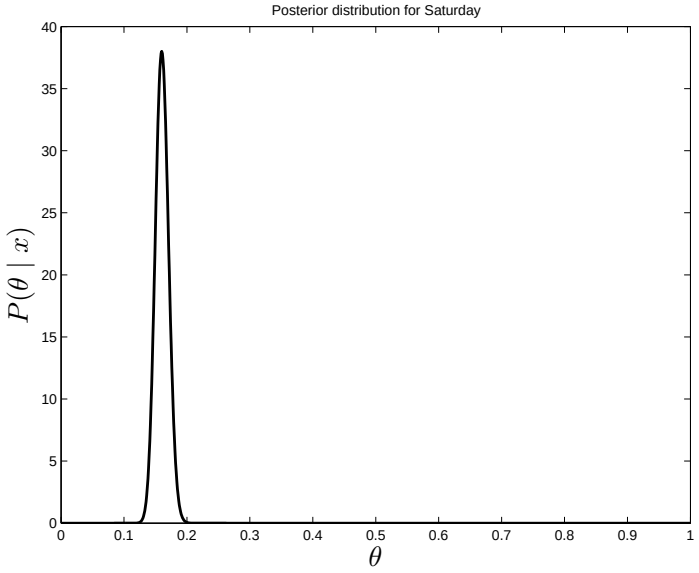




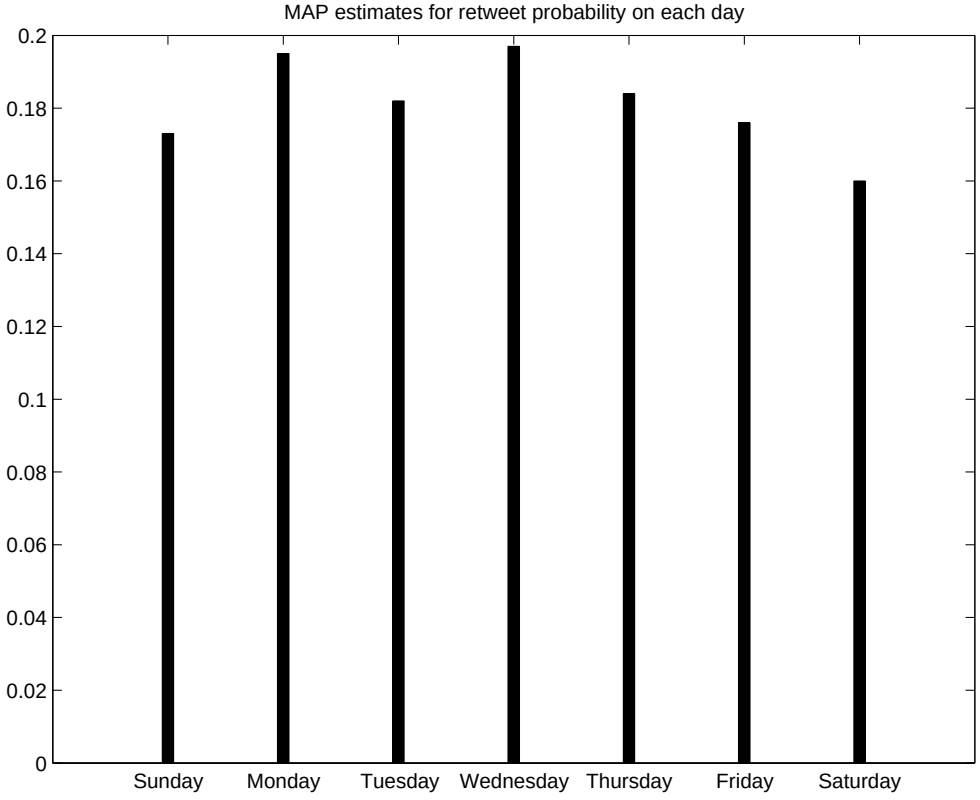








(d) Based on the posterior distribution does the day of the week on which a tweet is written have an effect on the chance of it being retweeted? Justify your answer.



1.5

Figure 1.2: The MAP estimates for each day incorporating all data points. $P(\theta; \alpha, \beta) = \text{Beta}(\theta; 2, 2)$ was used as the prior.

By examining the posteriors and the MAP estimates (Figure (1.2)) it is clear that the day of the week that the tweet is written does have an effect on its chance of being retweeted with the probability being highest on Wednesdays and lowest on Saturdays.

This is just based on mean analysis. It is also important to see if the posteriors overlap a lot

(e) Fred, a colleague, questions your choice of prior. What are your arguments about the validity of your work?

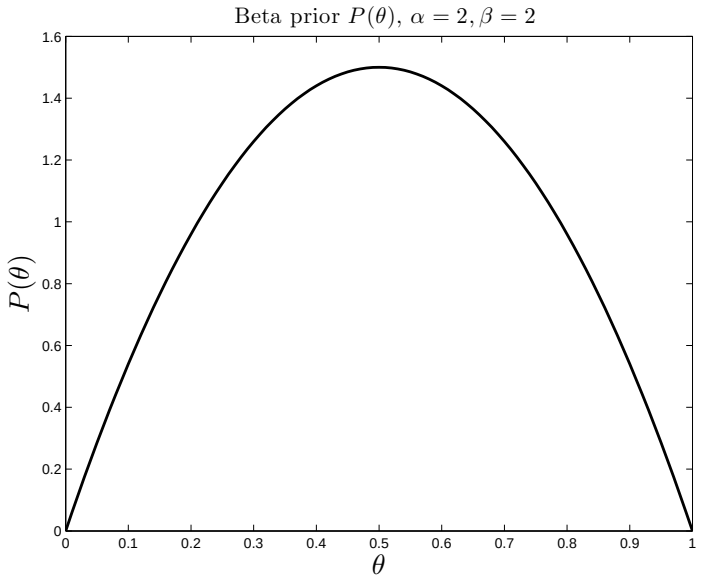


Figure 1.3: The Beta prior used for $P(\theta)$, $\alpha = 2$, $\beta = 2$.

This prior incorporates our beliefs that extreme values of θ are probably less likely and that we can never be certain that it will be retweeted or not (i.e. $P(0) = P(1) = 0$). The idea that (except for logical statements like $2 + 2 = 4$) probabilities should not be given the value of 0 or 1 is known as *Cromwell's Rule*[1] and essentially states that certainties should be avoided as even seemingly impossible events may have a very small possibility of occurring.

The use of $\alpha = \beta = 2$ essentially adds a pseudocount of 1 to the number of observed positive and negative cases. This prevents the probabilities of having a positive or negative case from being zero even if they are not observed in the training data and is known as *additive smoothing*, *Laplace smoothing*, or in the special case where the added pseudocount is one (which we have here), *add-one smoothing*.

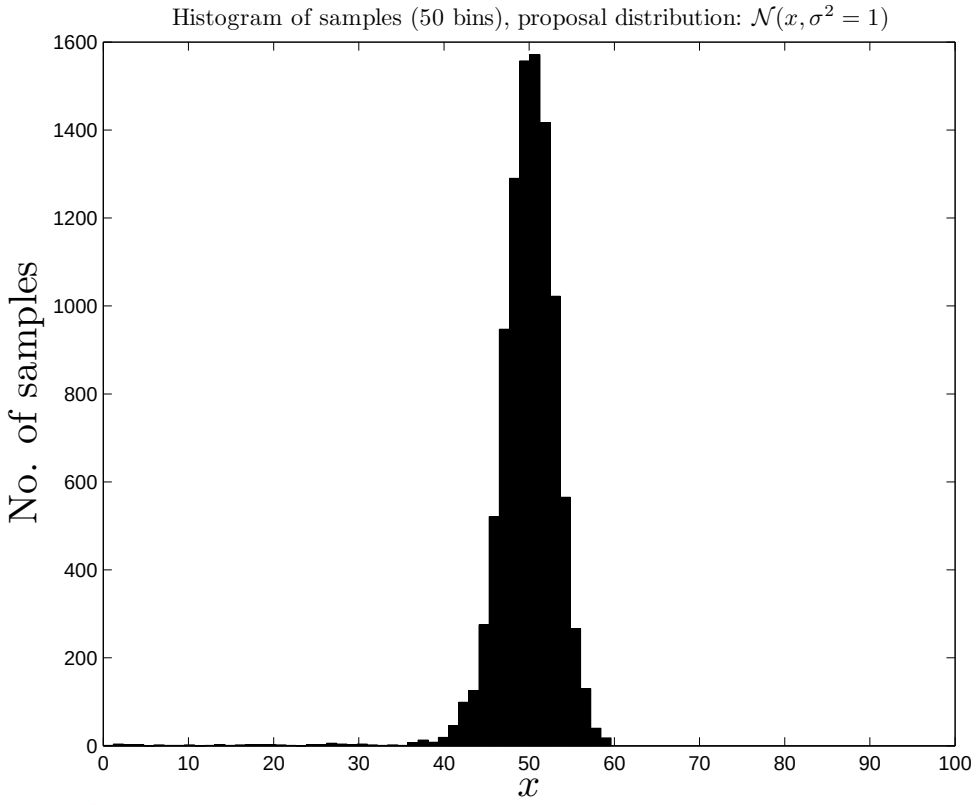
A uniform prior of $\alpha = \beta = 1$ on the other hand would mean that the probability of certain events would be zero if they were not observed in the training data. This is not desirable behaviour in the real world where the training data may not cover all possibilities.

The Beta distribution is used because it is a conjugate prior to the Bernoulli distribution of the likelihood and thus allows for the use of relatively simple analytical methods to obtain the posterior.

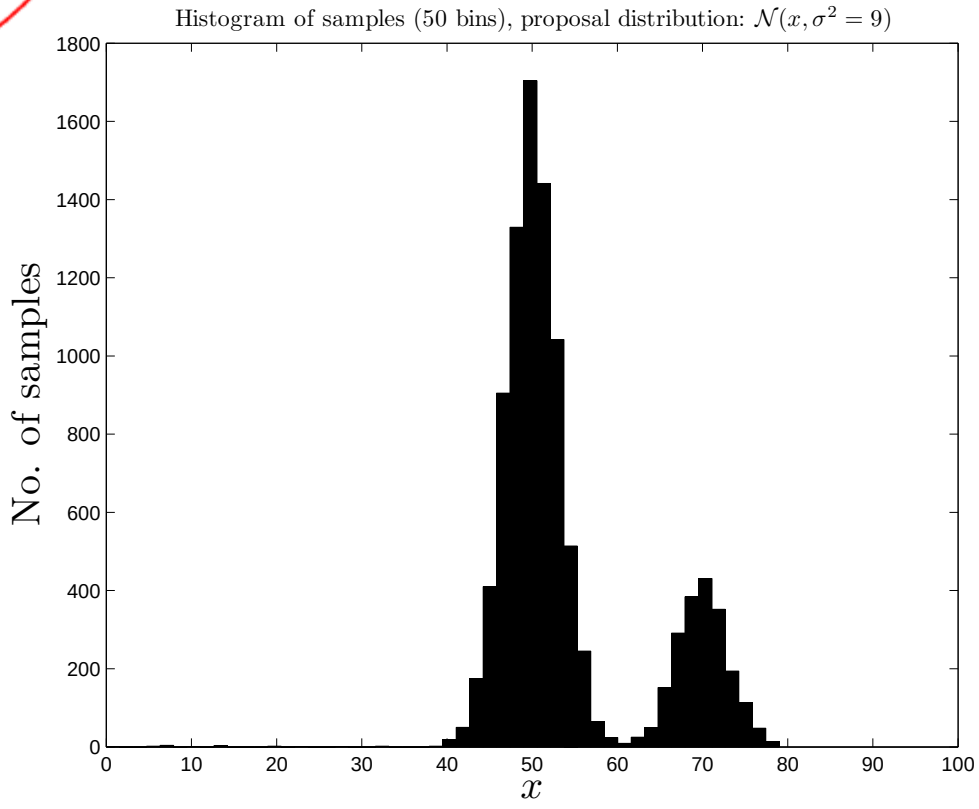
[1] Named after Oliver Cromwell, the leader of the Roundheads during the English Civil War, who wrote to the synod of the Church of Scotland: "*I beseech you, in the bowels of Christ, think it possible that you may be mistaken*"

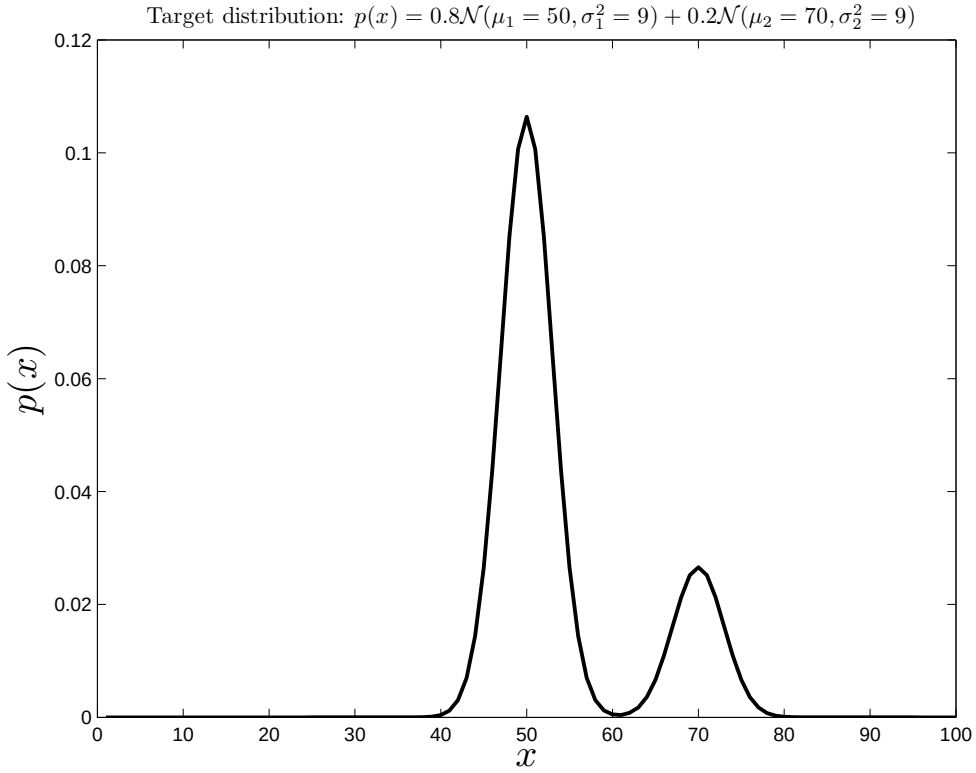
2 Question 2 *(15 marks)*

(a)...Run the sampler for 10,000 iterations, with an initial value x = 0, for both proposal distributions. Plot the two histograms for both sets of samples and the target distribution on three separate plots.



8





(b) Explain the difference between the two histograms. Based on this, explain which is the better proposal distribution and why.

The first proposal distribution is narrower than the second due to its smaller variance. This means that the sampler is unable to traverse the low-probability region between the two peaks and instead just stays at the first peak at $x = 50$ continuing to sample that whilst completely missing the secondary peak at $x = 70$.

The second proposal distribution is broader and therefore is able to traverse to the secondary peak and provide a reasonably accurate estimate of the target distribution. For the narrower proposal distribution 89.11% of proposed samples were accepted, this is a higher acceptance rate than one would desire as it means that the sampler is probably not sampling the full space of the target distribution sufficiently well and is just staying around highly probable regions.

For the broader proposal distribution the acceptance rate was 70.27% which is still higher than the general optimal value of 23.4% (although 50% may be good for a 1-D Gaussian distribution) but we can see from the histograms that it performed better and was indeed able to sample the full space of the target distribution, even if it did not do so in an optimal manner (i.e. it takes longer than necessary to traverse the space and converge on $p(X)$).

(c) Fred, your ever helpful colleague, points out that your algorithm produces too many sample points far away from the distribution and suggests you initialise x to the mean of the distribution x = 54.
(i) Explain whether this is a good approach.
(ii) What other approach could you take to deal with this problem.

In general this is not a good approach because for bimodal distributions the mean could be in a region of low probability and far from either peak, and if the target distribution is essentially flat there then it could take a long time to traverse out from that point.

An approach that could deal with this is *simulated annealing*. This involves slowly decreasing the probability of accepting less probable samples as the sampler explores the space (i.e. the higher the current iteration number, the less likely it is to accept a less probable sample). This means that initially the sampler moves quite rapidly through the space (thus enabling it to quickly get from the distant initialisation point to the more probable regions), but retains the ability to efficiently sample the distribution once it is there as it won't waste time moving to highly improbable states.

3 Question 3 *(25 marks)*

(a) Fred suggests that a Gibbs sampling approach could be used. Is he correct? Explain your reasoning.

In this case we have 58 features and so the problem has high-dimensionality. Therefore it becomes hard to avoid very slow convergence as the region of high-probability space is small compared to the total space and so if each step changed every weight, it would be unlikely to land there.

Gibbs' sampling solves this problem by sampling each variable separately and immediately using the updated value for subsequent samples.

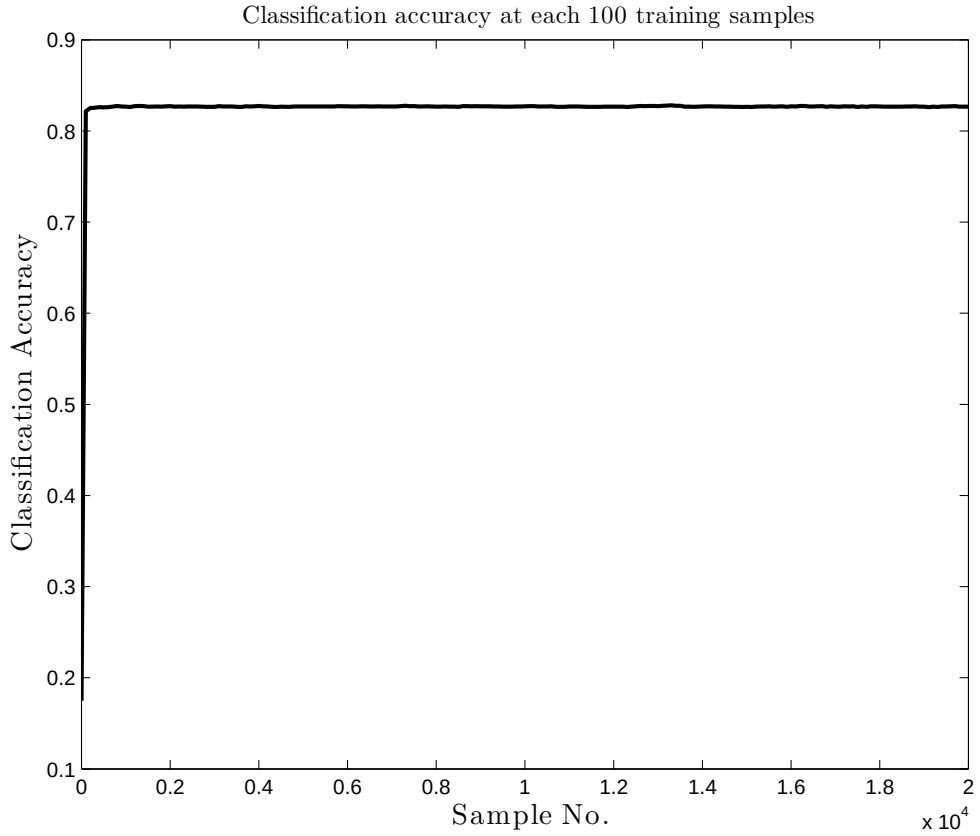
(b) Use your Metropolis-Hastings sampler to sample from the posterior. Start with all weights at 0 and run the sampler for 20,000 samples with no burn in.
(i) Describe your chosen proposal distribution. Provide evidence that the resulting sampler has acceptable performance.

I chose to use a multivariate Gaussian distribution with the identity matrix as the covariance matrix as the proposal distribution i.e. $q(\mathbf{w}'|\mathbf{w}) = \mathcal{N}(\mathbf{w}, \mathbf{I})$

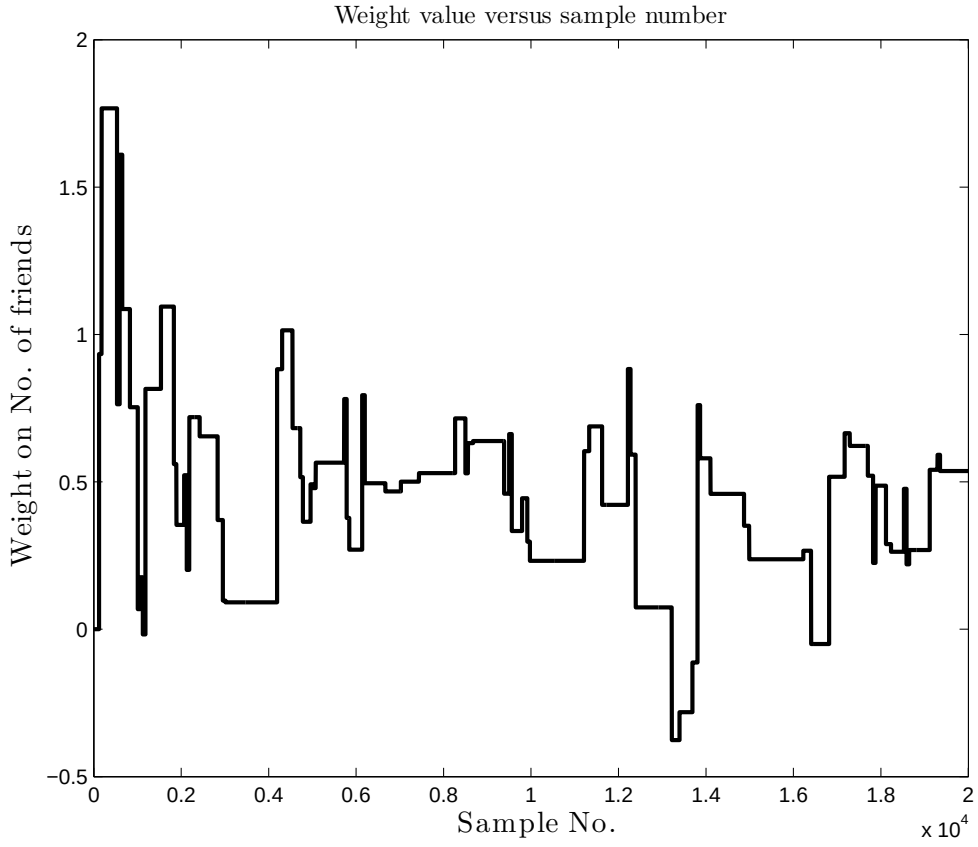
The resulting sampler had an acceptance rate of 13.00% which is reasonably close to the optimal value of 23.4% so the performance is acceptable.

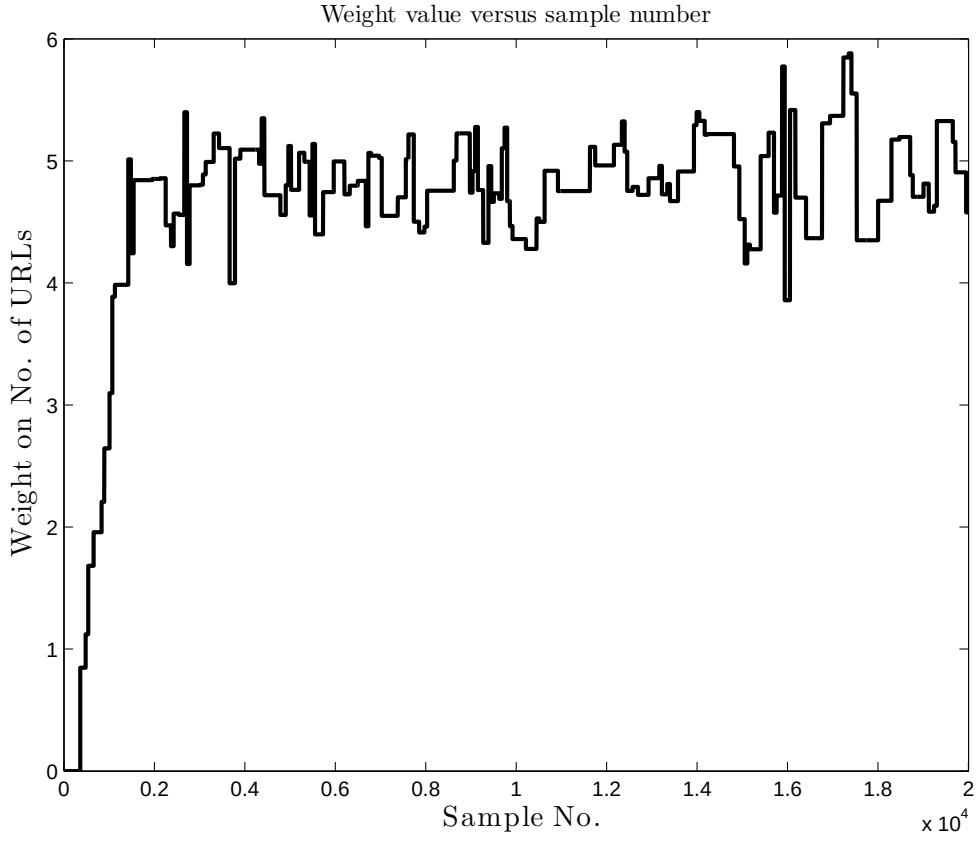
(ii) Plot the classification accuracy (i.e. the fraction of correct classifications on the test data set) for every 100 training samples.

Although the accuracy has some use as a figure of merit it is worth remembering that the test-set only contained 17.49% positives and therefore a classifier which simply predicted 0 for all tweets would still attain 82.51% classification accuracy. This is one reason why the Area Under the Curve (AUC) value of the Reciever Operator Characteristic curve (ROC curve) is generally a preferred figure of merit for classifier performance.

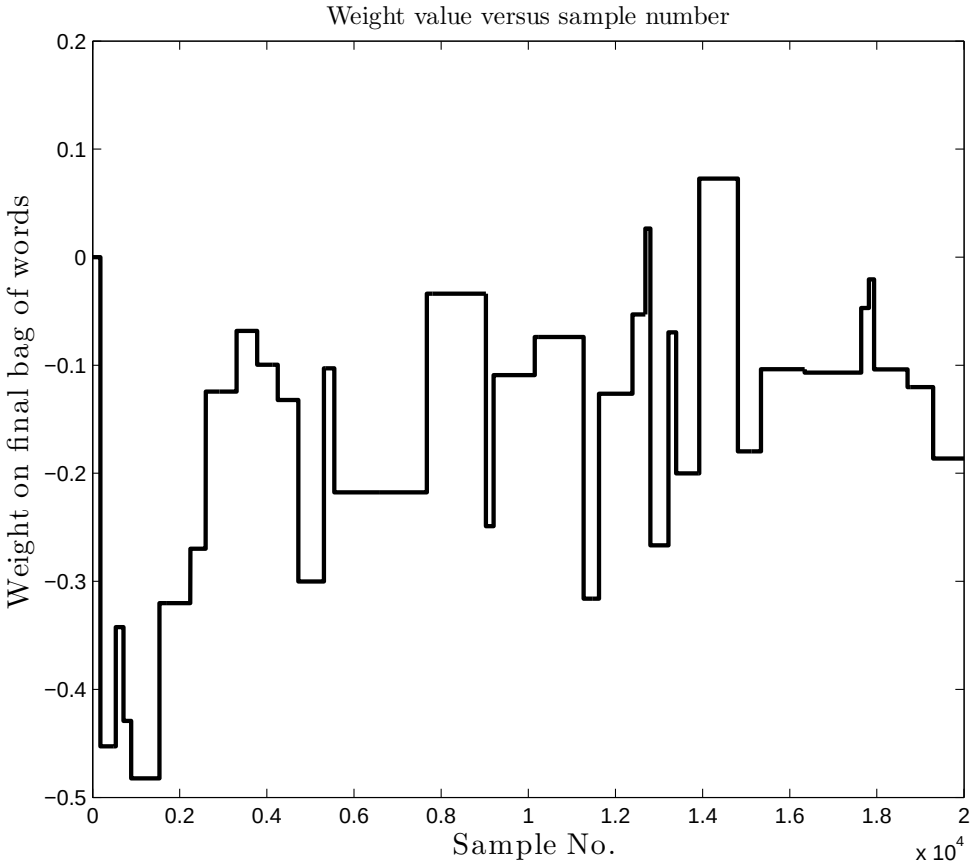


(iii) Select three weights and plot their values across the 20,000 samples.





(iv) Select the number of burn in samples. Explain your reasoning.

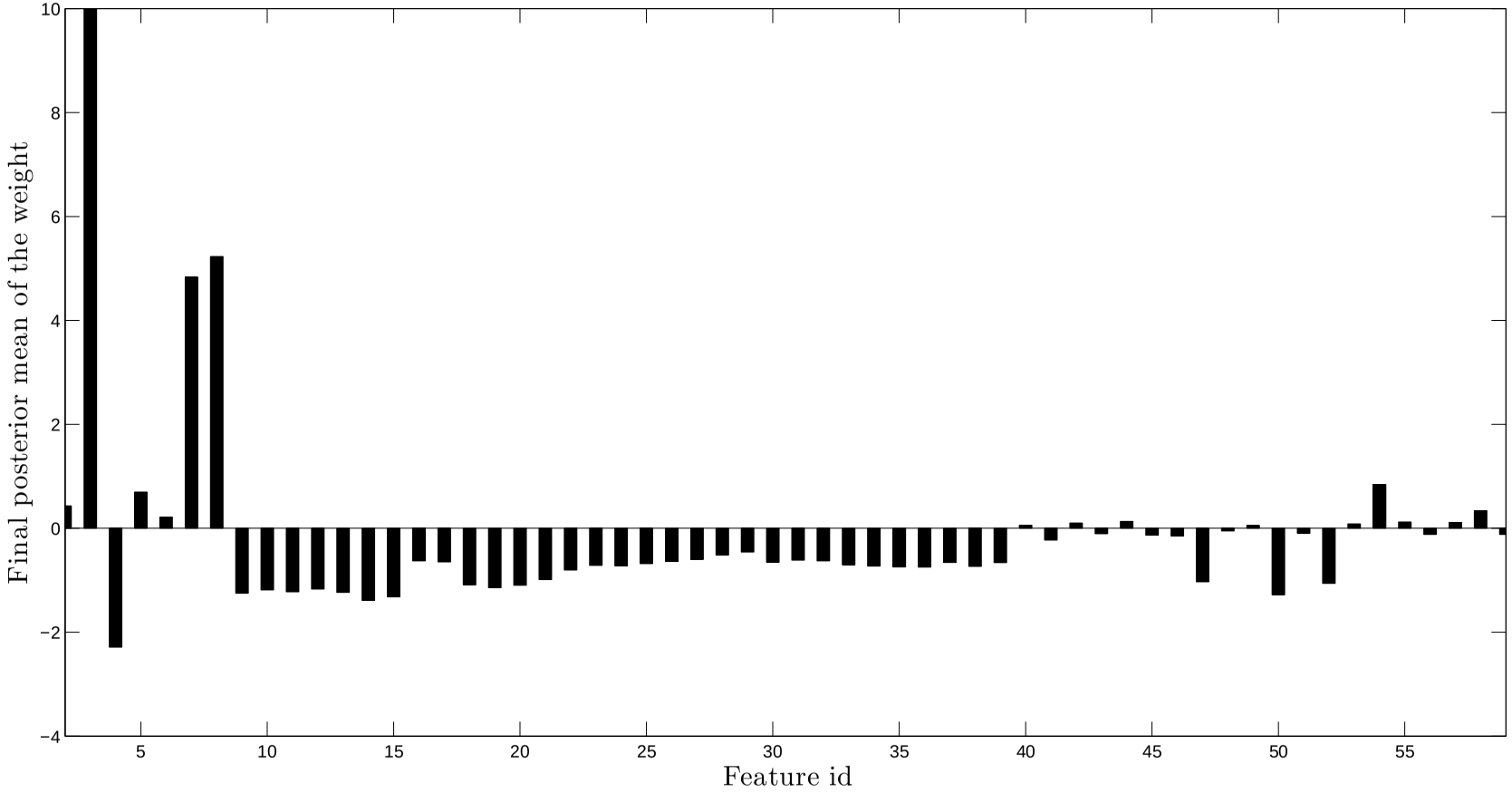


It seems that the values for the weights seem to settle down (except for fluctuations due to the probabilistic nature of the sampler) after 2000 samples, but to be sure I will choose the burn-in to be 5000 samples.

5

(c) Re-run the sampler for 20,000 samples, starting the weights at 0 but this time including your selected burn in.
(i) Plot the final posterior mean of the weights. (x-axis: feature id, using the column number in the features.mat file; y-axis: posterior mean.)

See following page.

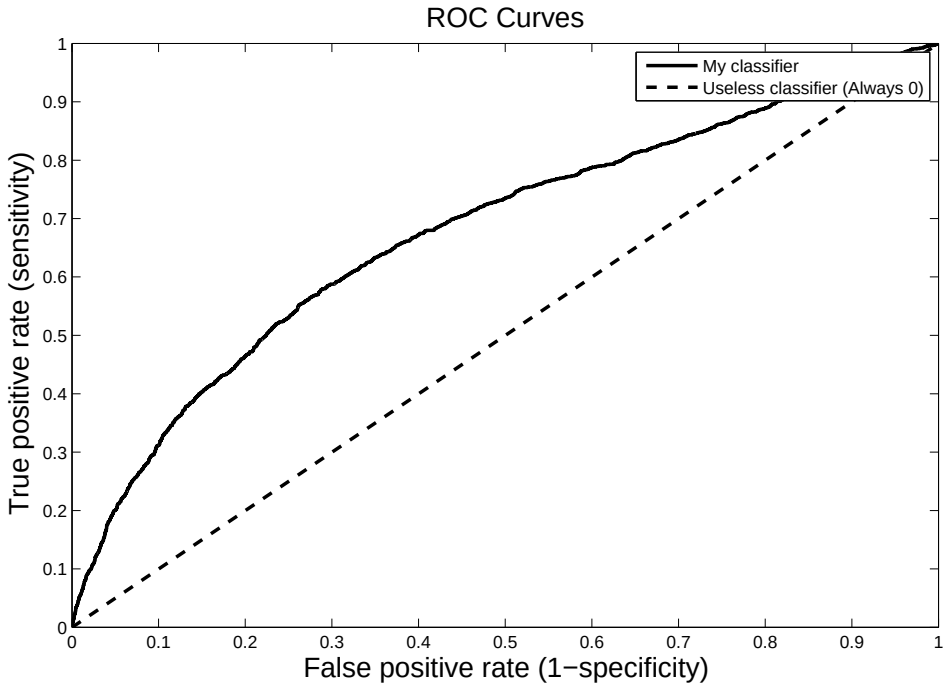

Final posterior mean of the weight
Feature id
10
8
6
4
2
0
−2
−4
5
10
15
20
25
30
35
40
45
50
55

(ii) Provide an interpretation of the weights and their relative importance in classifying whether a tweet will be retweeted.)

The features with the largest absolute weight values are the most important in classifying whether or not it will be retweeted. Features with negative weights make it less likely to be retweeted, features with positive weights make it more likely to be retweeted. It should be noted that as the weights are in the exponent in the sigmoid function, their direct relation to the predicted variable is not intuitive unlike for linear regression where they are the change in predicted variable per unit increase of the feature.

So we can see that the number of followers, number of mentions and number of hash tags make the tweet more likely to be retweeted. Whilst a high number of statuses makes the tweet less likely to be retweeted.

For completeness and to satisfy my curiousity about the performance of the classifier I calculated the Area Under the Curve (AUC) value from the Reciever Operator Characteristic curve (ROC curve) for my classifier. I found that the AUC value for my classifier (using the final posterior mean value of the weights) was 0.6779, the AUC value for a useless classifier (e.g. always 0) is 0.5.



Therefore we see that the classifier is working better than simply always guessing zero or randomly guessing, but it is not perhaps a brilliant classifier (ideally we would want the AUC value to be close to 1). This is probably because it is a linear classifier and therefore works best when the categories are close to being linearly separable. As we are dealing with real-world data with a high number of dimensions, this is unlikely to be the case.

Using a non-linear classifier such as an artificial neural network or support vector machines (via the kernel trick) may be more successful, but also more complex and can be far harder to interpret. By this I mean that although you would be able to more accurately predict whether a tweet would be retweeted given the features, you might not be able to tell why i.e. which features are the most important etc. easily.