Machine Learning and Pattern Recognition, Assignment Sheet 2013

School of Informatics, University of Edinburgh

Instructor: Charles Sutton

Handed out: Thursday 24^{th} October 2013
Submission Deadline: 4pm, Friday 15^{th} November 2013

This assignment is worth 20% of the total course marks.

To submit, convert your answers to `pdf` format. Check the file is readable in adobe reader on DICE, rename the file to `answers.pdf` and put it in directory called `answers`, along with any other files you used *including all code*, and type

```
submit mlpr 1 answers
```

Please ensure the `answers.pdf` file is self contained (don't put answers in your code). Please ensure that the answers you give are written in your own words and are explaining your understanding.

The marking will follow the standard University marking scheme. In the context of this assignment that means:

A Well explained description of points above plus extra achievement at understanding or analysis of results. Clear explanations, evidence of creative or deeper thought will contribute to a higher grade.

B Well explained answers to the questions.

C Fairly accurate answers to many questions, but significant deficiencies.

D Evidence that the student has gained some understanding, but not addressed that specified task properly.

E/F/G serious error or slack work.

Introduction

In this assignment you will be asked to examine data collated from tweets sent in May 2011. The tweets have already been pre-processed and features extracted (described in section B).

You *must* implement that required MATLAB scripts using base functionality only. In particular, you *must* not use any toolboxes.

1 Questions

1.1 Question 1 (10 marks)

The file `dowfeatures.mat` contains a small subset of features generated from user tweets. The first column indicates whether the tweet was re-tweeted. The other seven columns indicate the day of the week that the tweet was written (Sunday, Monday, ..., Saturday). All values are in the set $0, 1$.

The task is to use a Bayesian approach to determine whether tweets are more likely to be retweeted on certain days of the week.

(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.

   (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.

(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?

(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.

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

1.2 Question 2 (15 marks)

In this question you will create a Metropolis-Hastings sampler in MATLAB and test it on a simple distribution. This sampler will then be used in the following question. The distribution being approximated is given by

$$p(x) = 0.8\mathcal{N}(\mu_1 = 50, \sigma_1^2 = 9) + 0.2\mathcal{N}(\mu_2 = 70, \sigma_2^2 = 9) \tag{1}$$

IMPORTANT The MATLAB file for the sampler *must* be called `q2_sampler.m`. You *must* initialise the random seed to 2, which can be done by including the following at the beginning of your MATLAB script.

```
rand('state',2);
randn('state',2);
```

(a) Write a Metropolis-Hastings sampler in MATLAB to reproduce the distribution in Equation (1). You should implement and run the sampler with the two different proposal distributions:

$$q_1(x^*|x) = \mathcal{N}(x, \sigma^2 = 1) \tag{2}$$
$$q_2(x^*|x) = \mathcal{N}(x, \sigma^2 = 9) \tag{3}$$

where x is the current state of the Metropolis-Hastings sampler and x^* is the proposed new state.

Run the sampler for 10,000 iterations, with an initial value $x = 0$, for both proposal distributions in Equations (2) and (3). Plot the two histograms for both sets of samples and the target distribution in Equation (1) on three separate plots.

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

(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.

1.3 Question 3 (25 marks)

The file `features.mat` contains a larger set of features generated from user tweets. As before the first column indicates whether the tweet was re-tweeted. The remaining columns contain feature values based on both user and tweet information (see section B for details).

The goal is to create a linear classifier that can be used to determine whether a tweet will be retweeted based on the feature values, using an MCMC approach to fit the weights $\mathbf{w}$. The linear classifier will use a sigmoid function to determine the likelihood

$$p(y = 1|\mathbf{x}, w_0, \mathbf{w}) = \sigma(w_0 + \mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-(w_0+\mathbf{w}^T\mathbf{x})}} \tag{4}$$

where $y \in \{0, 1\}$ is the class of a single tweet (0 indicates not re-tweeted) and $\mathbf{x}$ are the extracted features of the tweet. Note that the number of data points is large and calculating the above probability distribution will result in numerical problems. The solution is to calculate the log probability, described in section A.

The features should be divided into a training data set and test data set using

```
SPLIT_INDEX = floor(size(features,1) * (4 /5));
traindata   = features(1:SPLIT_INDEX,:);
testdata    = features((SPLIT_INDEX + 1):end,:);
```

Your goal is to sample from the posterior distribution $p(w_0, \mathbf{w}|\mathcal{D})$ where

$\mathcal{D} = \{(y_1, \mathbf{x_1}), (y_2, \mathbf{x_2}), ...(y_N, \mathbf{x_N})\}$

is the full data set and the prior on the weights is $p(w_0, \mathbf{w}) = \mathcal{N}(\mathbf{0}, I)$.

IMPORTANT The MATLAB file for the sampler *must* be called `q3_sampler.m`. You *must* initialise the random seed to 2, which can be done by including the following at the beginning of your MATLAB script.

```
rand('state',2);
randn('state',2);
```

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

(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.

Hint: *Due to the high dimensionality, the region in the weight space that has high probability $p(w_0, \mathbf{w}|\mathcal{D})$ is relatively small. Use a proposal distribution that does not modify all the weights in each step.*

(i) Describe your chosen proposal distribution. Provide evidence that the resulting sampler has acceptable performance.

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

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

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

(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.)

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

A Log Probability

Using a simplistic calculation of the probability of the data for a given set of weights

$$p(\mathcal{D}|w_0, \mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x_i}, w_0, \mathbf{w}) \tag{5}$$

$$= \prod_{i=1}^{N} \sigma(w_0 + \mathbf{w}^T\mathbf{x})^{y_i}(1 - \sigma(w_0 + \mathbf{w}^T\mathbf{x}))^{(1-y_i)} \tag{6}$$

will result in an extremely small value that cannot be represented by MATLAB . To overcome this numerical problem we can use the log probability

$$\log p(\mathcal{D}|w_0, \mathbf{w}) = \sum_{i=1}^{N} \log\left[\sigma(w_0 + \mathbf{w}^T\mathbf{x})^{y_i}(1 - \sigma(w_0 + \mathbf{w}^T\mathbf{x}))^{(1-y_i)}\right] \tag{7}$$

$$= \sum_{i=1}^{N} \left[y_i \log(\sigma(w_0 + \mathbf{w}^T\mathbf{x})) + (1 - y_i)\log(1 - \sigma(w_0 + \mathbf{w}^T\mathbf{x}))\right] \tag{8}$$

Setting $\mathbf{w}_+ = (w_0, \mathbf{w})$, this modifies the Metropolis-Hastings acceptance calculation from

$$U[0, 1] \leq p(accept) = \min\left(1, \frac{p(\mathbf{w}_+{}^*|\mathcal{D})q(\mathbf{w}_+|\mathbf{w}_+{}^*)}{p(\mathbf{w}_+|\mathcal{D})q(\mathbf{w}_+{}^*|\mathbf{w}_+)}\right) \tag{9}$$

to

$$\log(U[0, 1]) \leq \min\left(0, \log(p(\mathbf{w}_+{}^*|\mathcal{D})q(\mathbf{w}_+|\mathbf{w}_+{}^*)) - \log(p(\mathbf{w}_+|\mathcal{D})q(\mathbf{w}_+{}^*|\mathbf{w}_+))\right) \tag{10}$$

Note that in the case where $q(\mathbf{w}_+{}^*|\mathbf{w}_+) = q(\mathbf{w}_+|\mathbf{w}_+{}^*)$ Equation (10) becomes

$$\log(U[0, 1]) \leq \min(0, \log(p(\mathbf{w}_+{}^*|\mathcal{D})) - \log(p(\mathbf{w}_+|\mathcal{D}))) \tag{11}$$

B Full Tweet Features (for Question 3)

The features are divided into sections, user related (columns 2-5) and tweet related (columns 6-59). The full set of features and their types (real or binary) are described in the table below and are found in the file `features.mat`. Real features have been normalised to contain values in the range $[0, 1]$. Binary features are in the set $0, 1$.

Feature	Column	Type
Is Retweeted?	1	Binary
User Features		
N Friends	2	Real
N Followers	3	Real
N Statuses	4	Real
N Favourites	5	Real
Tweet Features		
N URLS	6	Real
N Hash Tags	7	Real
N Mentions	8	Real
Day of the Week *(Sunday - 9)*	9-15	Binary
Hour of the Day *(Midnight - 16)*	16-39	Binary
Bag of Words	40-59	Binary