CSC 412/2506
Probabilistic Graphical Models
Assignment 1
Out: Jan. 19
Due: Feb. 9

1 Bayesian regression [20 points]

1. Consider $f = \mathbf{w}^T\mathbf{x}$, where $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \Sigma)$. Show that $p(f|\mathbf{x})$ is Gaussian.
2. Find the mean and covariance of f.
3. Consider the target $t = f + \epsilon$ where $\epsilon \sim \mathcal{N}(0, .001)$. What is $p(f|t, \mathbf{x})$?

2 Conjugate priors and exponential family [20 points]

Based on a question in Murphy.

Derive the conjugate prior for μ and $\lambda = 1/\sigma^2$ for a univariate Gaussian using the exponential family (see Murphy 9.2.5.5). By suitable reparametrization, show that the prior has the form $p(\mu, \lambda) = \mathcal{N}(\mu|\gamma, (\lambda\tau)^{-1})\text{Gamma}(\lambda|\alpha, \beta)$.

Note that it is possible to derive a 3-parameter variation instead of the 4-parameter version above. This will take the form $p(\mu, \lambda) = \mathcal{N}(\mu|\gamma, (\lambda(2\alpha - 1))^{-1})\text{Gamma}(\lambda|\alpha, \beta)$. Either variant is fine.

3 Spam classification using logistic regression [30 points]

From Hastie, Murphy.

There are files spambase.train.txt and spambase.test.txt (available on the course website) in which each row is an email message. The final column in each file indicates whether the email was spam. 57 features have been extracted from each email message. These are as follows:

- 48 features in [0, 100], giving the percentage of words in a given message which match a given word on a list containing words such as “business”, “free”, “george”, etc.
- 6 features in [0, 100], giving the percentage of characters in the email that match these characters: ; ([! $ #

- Feature 55: The average length of an uninterrupted sequence of capital letters.
- Feature 56: The length of the longest uninterrupted sequence of capital letters.
- Feature 57: The sum of the lengths of uninterrupted sequences of capital letters.

The dataset consists of 4601 email messages, which we've divided into 3000 training and 1601 test examples.

1. Fit a ℓ_2-regularized logistic regression model to the data. Use 5-fold cross-validation to determine an appropriate regularization coefficient. Report the mean error rate on the training and test sets you obtain, as well as the value of the regularization coefficient.
2. An important decision concerns whether and how to pre-process the data. One typical choice is to standardize each input feature so that it has mean zero and variance one. Do this and evaluate the model again, and compare to your previous results.
3. Another popular choice is to represent data based on their logs. Transform the features using $\log(x_{ij} + 1)$, and then retrain the model and compare the results to your earlier ones.
4. In some cases it is sufficient to just binarize the data. Convert each feature to 1 iff its value is non-zero, and then retrain and compare the model.
5. Using the model that achieves the lowest *test* error, which 5 features are most informative in determining that an email should be marked as spam? Which 5 features are most informative in determining that an email should be marked as not spam?

Note: you may use publicly available software to complete this question, as long as you cite it in your writeup.

4 Collaborative filtering data [30 points]

Download the popular collaborative filtering dataset MovieLens-100k from `http://grouplens.org/datasets/movielens/`. The dataset consists of 100000 movie ratings from 943 users and 1682 movies. The ratings range from 1 to 5, with 5 being the best rating. This question only requires the ratings, which can be found in the third column of the u.data file (see the README for more information on this dataset).

1. Examine and plot the empirical marginal distribution of ratings. Can you explain why this distribution has the form that it does? Consider how the ratings are collected.

2. Fit a Gaussian to the ratings distribution using maximum likelihood. Describe this procedure and justify whether you think this is a good fit.

3. The beta-binomial distribution (see Murphy 3.3, or http://en.wikipedia.org/wiki/Beta-binomial_distribution) is derived by combining a binomial likelihood on the ratings with a beta prior on the parameter p of the binomial distribution. Assuming there are N ratings, write the log-likelihood of the beta-binomial distribution in the form $\sum_{i=1}^{N} \log \mathcal{L}(x_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the parameters of the distribution, $\mathcal{L}$ represents the likelihood, and x_i represents a rating. Note that the domain of the beta-binomial distribution is $\{0, \ldots, r-1\}$, where r is the maximum rating, so you will have to subtract 1 from each rating. Fit the beta-binomial distribution using either the method of moments, or by using numerical optimization to maximize the log-likelihood. Report the parameters that you find. Justify whether you think the beta-binomial is a good fit to this data.

4. Divide the data into training and test. Compare the log-probabilities of of the test data based on fitting these two distributions to the training data. Which model is more suitable for this data?

Please compare the distributions in this question qualitatively. Since one is continuous while the other is discrete, a quantitative comparison is complicated.

Note: you may use publicly available software to complete this question, as long as you cite it in your writeup.

Submission Instructions

You should submit a single pdf file (**hw1-your-student-id.pdf**) that contains your written solution to questions 1 and 2, and a report on the next two questions. These reports should answer briefly the questions we asked (in a few sentences) and also include some plots when they are useful, such as of the marginal ratings distribution with a fitted distribution shown on top of it. The entire file should not be more than 10 pages long, preferably 6-7 pages including plots. You do not need to turn in any code.

Your pdf should be submitted to MarkUs (https://markus.cdf.toronto.edu/csc412-2016-01) by 3pm on February 9th. You should be able to login using your CDF accounts. Penalties will be assessed for any assignment turned in later than this. **Please make sure that the submitted file is named hw1-your-student-id.pdf**.