CPSC-540 Probabilistic Machine Learning 2011

Homework # 2

Due Thursday, Th Feb 3rd 12:30pm.

NAME:_____

Signature:_____

STD. NUM: _____

General guidelines for homeworks:

You are encouraged to discuss the problems with others in the class, but all write-ups are to be done on your own.

Homework grades will be based not only on getting the "correct answer," but also on good writing style and clear presentation of your solution. It is your responsibility to make sure that the graders can easily follow your line of reasoning.

Try every problem. Even if you can't solve the problem, you will receive partial credit for explaining why you got stuck on a promising line of attack. More importantly, you will get valuable feedback that will help you learn the material.

Please acknowledge the people with whom you discussed the problems and what sources you used to help you solve the problem (e.g. books from the library). This won't affect your grade but is important as academic honesty.

When dealing with python exercises, please attach a printout with all your code and show your results clearly.

1 MLE for the uniform distribution

Consider a uniform distribution centered on 0 with width $2a$. The density function is given by

$$p(x) = \frac{1}{2a} I(x \in [-a, a]) \tag{1}$$

1. Given a data set $x_1, \ldots, x_n$, what is the maximum likelihood estimate of a (call it $\hat{a}$)?

2. What probability would the model assign to a new data point x_{n+1} using $\hat{a}$?

3. Do you see any problem with the above approach? Briefly suggest (in words) a better approach.

2 MLE for the Poisson distribution

The Poisson pmf is defined as $\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \ldots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

3 Gradient and Hessian of log-likelihood for logistic regression

1. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a)) \tag{2}$$

2. Using the previous result and the chain rule of calculus, derive an expression for the log-likelihood of binary logistic regression, as presented in class.

3. The Hessian can be written as $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$, where $\mathbf{S} := \mathrm{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Show that $\mathbf{H}$ is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of $\mathbf{S}$ will be strictly positive, and that $\mathbf{X}$ is full rank.)

4 Gradient and Hessian of log-likelihood for multinomial logistic regression

1. Let $\mu_{ik} = \mathcal{S}(\boldsymbol{\eta}_i)_k$. Prove that the Jacobian of the softmax is

$$\frac{\partial \mu_{ik}}{\partial \eta_{ij}} = \mu_{ik}(\delta_{kj} - \mu_{ij}) \tag{3}$$

where $\delta_{kj} = I(k = j)$.

2. Hence show that

$$\nabla_{\mathbf{w}_c} \ell \quad = \quad \sum_i (y_{ic} - \mu_{ic}) \mathbf{x}_i \tag{4}$$

Hint: use the chain rule and the fact that $\sum_c y_{ic} = 1$.

3. Show that the block submatrix of the Hessian for classes c and c' is given by

$$\mathbf{H}_{c,c'} \quad = \quad -\sum_i \mu_{ic} (\delta_{c,c'} - \mu_{i,c'}) \mathbf{x}_i \mathbf{x}_i^T \tag{5}$$

5 Logistic regression in Matlab

Download the dataset `tremor.mat` from the course website. A description of this dataset appears in:

http://www.mitpressjournals.org/doi/pdf/10.1162/089976601750541831

This is a two-class classification problem with two-dimensional input features. You can load the data into matlab as follows:

```
load tremor;
data =[x_tr t_tr];
[N,arb] = size(data);         % N= Number of data points.
data = data(randperm(N),:);   % Order the data randomly.
xv = x_te';                   % Test set input data.
dv = t_te;                    % Test set target data.
x = data(:,1:2)';             % Train set input data.
d = data(:,3);                % Train set target data.
```

Just as we did in class for the XOR classification problem with logistic regression, your task is to fit linear logistic regression, quadratic logistic regression, and kernel logistic regression to the tremor training data. For kernel logistic regression, place the kernel centers at the data points. You can use Gaussian kernels, thin-plate spline kernels, logistic kernels or any other kernel of your choice. Please hand in 3 Figures and the code. Each figure should show the training and test points and the decision boundary. You should also compute the percentage of classification errors for each method:

```
percentageError = sum(abs(yp-d))*100/N  % Train error where the prediciton yp is 0 or 1.
```

The person that gets the lowest test set error gets an extra 2 marks in the midterm. Careful with over-fitting the test set!