Machine Learning 10-701 March 18, 2013
http://alex.smola.org/teaching/cmu2013-10-701/ due on April 1, 2013
Carnegie Mellon University

Homework 4

Instructions

- The homework is due in the lecture on April 1. Any homework received after the lecture will not be considered.
- Please submit one set of notes for each of the problems and put them into a separate stack. Don't forget to add your name on each sheet.
- Alternatively, you can e-mail your solution to `10.701.homework@gmail.com`. Please write the subject as: `[Homework 3] yourandrewID`, followed by the numbers of the questions you are including in the email, or `all` if you're submitting the entire homework. As before, the cutoff is the end of the lecture.
- If you submit code, it should be sufficiently well documented that the TAs can understand what is happening. Also attach pseudocode if you feel that this makes the result more comprehensible.

Machine Learning 10-701 March 18, 2013
http://alex.smola.org/teaching/cmu2013-10-701/ due on April 1, 2013
Carnegie Mellon University

Homework 4

---

1 Rating Distribution [25 points]

Your goal is to model product rating distributions. Many websites, such as Yelp provide this kind of information. In many cases one simply uses a least-mean-squares loss to fit the data. However, this isn't terribly accurate, when looking at real distributions, such as those on yelp.com:

- BRGR (unimodal) http://www.yelp.com/biz/brgr-pittsburgh
- Legume (skewed) http://www.yelp.com/biz/legume-pittsburgh
- Burma Tokyo (bimodal) http://www.yelp.com/biz/burma-tokyo-pittsburgh

Likewise, the reviews on newegg.com are not necessarily unimodal:

- Crucial SSD (bimodal) http://www.newegg.com/Product/Product.aspx?Item=N82E16820148444&SortFi
  0&SummaryType=0&PageSize=100&SelectedRating=-1&VideoOnlyMark=False&IsFeedbackTab=
  true#scrollFullInfo
- Samsung SSD (skewed) http://www.newegg.com/Product/Product.aspx?Item=N82E16820147187&SortF
  0&SummaryType=0&PageSize=100&SelectedRating=-1&VideoOnlyMark=False&IsFeedbackTab=
  true#scrollFullInfo

It is your goal to design an exponential family distribution that can be used to estimate ratings.

1.1 Exponential Family

For ratings $x \in \{1 \dots R\}$ use the statistic

$$\phi(x) = (x, x^2) \tag{1}$$

to design an exponential family distribution over ratings (an algebraic expression is OK).

1. Explain how to model unimodal distributions using (1).
2. Explain how to model bimodal distributions using (1).
3. Explain how to model skewed distributions using (1).

1.2 Log-Partition Function

1. Write out the log-partition function $g(\theta)$.
2. Compute the variance of x using *only* the first derivative of $g(\theta)$.

1.3 Parameter Inference

For the following empirical rating distributions compute optimal values of θ and provide plots of the estimated rating distributions $p(x|\theta)$. Hint - use an off-the-shelf convex function minimizer, e.g. in Octave or Matlab. The ratings below are all taken from amazon.com as of March 18, 2013.

	1	2	3	4	5
Panasonic LX-7	3	1	13	32	80
Seagate 4TB disk	81	19	37	120	255
Machine Learning (Tom Mitchell)	1	2	5	8	28
Machine Learning for Hackers (Drew Conway)	0	5	8	3	4
Blackberry 9650 Bold	80	14	21	22	81

2 Exponential Families [25 points] (Mu)

2.1 Entropy of exponential families

Compute the entropy of an exponential family distribution

$$p(x;\theta) = \exp\left(\langle \phi(x), \theta \rangle - g(\theta)\right) \tag{2}$$

in terms of θ, $g(\theta)$ and $\partial_\theta g(\theta)$.

2.2 Multivariate Gaussian

For vectors $x \in \mathbb{R}^k$ the density function of k-dimensional Gaussian distribution is given by

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right). \tag{3}$$

Show that its entropy is given by

$$H(p) = \frac{1}{2}\ln|\Sigma| + \frac{k}{2}(1+\ln(2\pi)) \tag{4}$$

Hint — to show this you can either use the result of the previous question or compute it directly.

2.3 Conjugate Distribution

The Gamma distribution is defined as

$$\text{Gamma}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1}\exp(-b\lambda). \tag{5}$$

Show that for a univariate Gaussian distribution $\mathcal{N}(x|\mu,\sigma^2)$ the conjugate prior is the Gaussian-Gamma prior. Rather than looking up the derivation in Wikipedia follow these steps:

1. Rewrite the Normal distribution in terms of natural parameters θ and sufficient statistics $\phi(x)$.
2. Rewrite the Gamma distribution in terms of natural parameters θ, prior means μ_0 and counts m_0.
3. Write out the posterior probability distribution under this model.

Homework 4

---

3 Gaussian Process Regression [30 points] (Xuezhi)

In this question you will need to implement Gaussian Process Regression. You can use any programming language you like (preferably Matlab/R). Just invoking a toolbox is **not OK**. In your writeup please attach all the plots and answers to the questions. Also send your code to the submission email.

In Gaussian Process Regression, we see training data $\mathbf{x} = [x_1, ..., x_n]^\top$, and the corresponding labels $\mathbf{y} = [y_1, ..., y_n]^\top$. Now we receive test data points $\mathbf{x}^*$, and we want to predict $\mathbf{y}^*$. Assume in this question the mean $\mu = 0$. The joint distribution is

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_* \\ K_*^\top & K_{**} \end{bmatrix}\right) \tag{6}$$

From previous lectures we know that

$$y_* \sim \mathcal{N}(K_*^\top K^{-1} y, K_{**} - K_*^\top K^{-1} K_*) \tag{7}$$

where the kernel generating the covariance matrices is given by $\bar{k}(x, x') = k(x, x') + \sigma^2 \delta(x, x')$. In your experiments you are going to use the Gaussian RBF kernel

$$k(x, x') = \exp\left(-\frac{1}{2\gamma^2} \|x - x'\|^2\right) \tag{8}$$

3.1 Regression Code

1. Generate training and test data using

```
sigma_n=0.5;
x      = [-5:1:5]';
y      = 2*sin(x) + sigma_n*randn(length(x), 1);
x_star = -5:0.1:5;
y_star = 2*sin(x_star);
```

2. Implement Gaussian Process Regression. Use bandwidth $\gamma^2 = 1$.
3. Plot the results (in the same graph using different colors):
   - Training data points (x-axis) and correspoding labels (y-axis).
   - Test data points, the true labels and the predicted labels.
   - Standard deviation of the predicted function.

3.2 Changing the Setting

1. Less Training Data: Repeat the above procedure with fewer tranining data points, compare the results and explain the changes.

```
x=[-5:2:5]';
```

2. Change of Bandwidth: Repeat the above procedure using $\gamma^2 = 10$ and as alternative bandwidths $\gamma^2 = 0.5$. As before, plot the graph and explain what has changed and why.
3. Change of Covariate Distribution: Repeat the procedure but now generate the training data nonuniformly. Report and explain the changes.

```
x=[-5:2:-1 0:0.1:5]';
```

4 Information Theory [20 points] (Junier)

1. For which probability is the entropy of Bernoulli random variables maximized?
2. Prove the conditional entropy decomposition

$$H[X, Y] = H[X] + H[Y|X] \tag{9}$$

3. Assume that we have a Markov chain with

$$p(x) = p(x_1) \prod_{i=1}^{n-1} p(x_{i+1}|x_i), \tag{10}$$

use (9) to derive an entropy decomposition for $p(x)$. Hint — you should obtain terms involving only (x_i, x_{i+1}).

4. Assume that we are given an initial distribution

$$p(x_0) = \begin{cases} 0.4 & \text{if } x_0 = 0 \\ 0.6 & \text{if } x_0 = 1 \end{cases} \tag{11}$$

and a transition matrix

$$\Pi = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix} \text{ where } \Pi_{ab} = p(x_{i+1} = a|x_i = b). \tag{12}$$

What is the entropy for a chain of length 3? Hints:

- Compute entropies for the conditional probabilities arising from Π.
- Take expectations with respect to $p(x_i)$.
- Use octave or Matlab to avoid doing matrix multiplication manually.