CPSC-540 Probabilistic Machine Learning 2011

Homework # 3

Due Thursday, Th March 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 Linear regression with kernels

The file `linregRbfDemo.m` in PMTK implements linear regression for Gaussian kernels. Instead of doing linear regression, I'd like you to try ridge regression with 3 kernel widths equal to 0.1, 0.2, and 0.6 (instead of the 3 current kernel widths). For your specific dataset and each value of the kernel width, estimate the best value of the regulariser either by 5-fold cross-validation or by computing the evidence. Hand in your code, plots of the fits for each of the 4 values of the kernel width at the optimum value of the regularizer that you found and at the value of the regularizer that the code currently uses. Hand in also a plot of the evidence or cross-validation error as a function of the regularization coefficient for the kernel width 0.2. In total, you should hand in 9 plots (or subplots).

2 MLE and MAP for multinoullis

Suppose $X \in \{1, 2\}$ and $Y \in \{1, 2, 3\}$. Define the joint distribution $P(X = j, Y = k) = \theta_{j,k}$. Consider the training data $\mathcal{D}$ below, where row i represents x_i and y_i:

X	Y
1	1
2	2
1	3
1	1
2	2
2	3

1. Find the maximum likelihood estimates. Hint: just build the joint 2×3 table of counts and normalize so all numbers sum to one.

2. Now suppose $p(\boldsymbol{\theta}) = \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_6)$, where $\alpha_\ell = 1$ for $\ell = 1:6$ (here we use ℓ to represent the double index (j,k), since we require $\sum_\ell \theta_\ell = 1$). What is the MAP estimate $\hat{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{\theta}) p(\mathcal{D}|\boldsymbol{\theta})$?

3 Mean, mode, variance for the beta distribution

Suppose $\theta \sim \mathrm{Beta}(a, b)$. Show that

$$\mathbb{E}[\theta] = \frac{a}{a+b} \tag{1}$$

$$\mathrm{var}[\boldsymbol{\theta}] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2}$$

$$\mathrm{mode}[\boldsymbol{\theta}] = \frac{a-1}{a+b-2} \tag{3}$$

4 Beta updating from censored likelihood

Suppose we toss a coin $n = 5$ times. Let X be the number of heads. We observe that there are fewer than 3 heads, but we don't know exactly how many. Let the prior probability of heads be $p(\theta) = Beta(\theta|1, 1)$. Compute the posterior $p(\theta|X < 3)$ up to normalization constants, i.e., derive an expression proportional to $p(\theta, X < 3)$.

5 Bayesian analysis of the Poisson distribution

In his book, Kevin defines the Poisson distribution with rate λ and derives its MLE. Here we perform a conjugate Bayesian analysis.

1. **5 points** Derive the posterior $p(\lambda|D)$ assuming a conjugate prior $p(\lambda) = \mathrm{Ga}(\lambda|a, b) \propto \lambda^{a-1}e^{-\lambda b}$. Hint: the posterior is also a Gamma distribution.

2. **5 points** What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a $\mathrm{Ga}(a, b)$ distribution is a/b.)

6 Posterior predictive distribution for a batch of data with the dirichlet-multinomial model

In class, we discussed the the posterior predictive distribution for a single multinomial trial using a dirichlet prior. Now consider predicting a *batch* of new data, $\tilde{\mathcal{D}} = (X_1, \ldots, X_m)$, consisting of m single multinomial trials (think of predicting the next m words in a sentence, assuming they are drawn iid). Derive an expression for

$$p(\tilde{\mathcal{D}}|\mathcal{D}, \alpha) \tag{4}$$

Your answer should be a function of $\boldsymbol{\alpha}$, and the old and new counts (sufficient statistics), defined as

$$N_k^{old} = \sum_{i \in \mathcal{D}} I(x_i = k) \tag{5}$$

$$N_k^{new} = \sum_{i \in \tilde{\mathcal{D}}} I(x_i = k) \tag{6}$$

Hint: recall that, for a vector of counts, $N_{1:K}$, the marginal likelihood (evidence) is given by

$$p(\mathcal{D}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{7}$$

where $\alpha = \sum_k \alpha_k$ and $N = \sum_k N_k$.

7 Bayesian linear regression in 1d with known σ^2

Consider fitting a model of the form

$$p(y|x, \boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{8}$$

to the data shown below:

```
x = [94,96,94,95,104,106,108,113,115,121,131];
y = [0.47, 0.75, 0.83, 0.98, 1.18, 1.29, 1.40, 1.60, 1.75, 1.90, 2.23];
```

1. Compute an unbiased estimate of σ^2 using

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{9}$$

   (The denominator is $N - 2$ since we have 2 inputs, namely the offset term and x.) Here $\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$, and $\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1)$ is the MLE.

2. Now assume the following prior on $\mathbf{w}$:

$$p(\mathbf{w}) = p(w_0)p(w_1) \tag{10}$$

   Use an (improper) uniform prior on w_0 and a $\mathcal{N}(0, 1)$ prior on w_1. Show that this can be written as a Gaussian prior of the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)$. What are $\mathbf{w}_0$ and $\mathbf{V}_0$?

3. Compute the marginal posterior of the slope, $p(w_1|\mathcal{D}, \sigma^2)$, where $\mathcal{D}$ is the data above, and σ^2 is the unbiased estimate computed above. What is $\mathbb{E}\left[w_1|\mathcal{D}, \sigma^2\right]$ and $\mathrm{var}\left[w_1|\mathcal{D}, \sigma^2\right]$ Show your work. (You can use Matlab if you like.) Hint: the posterior variance is a very small number!

4. What is a 95% credible interval for w_1?