CSE 515T (Spring 2015) Assignment 2

Due Monday, 16 February 2015

1. (Curse of dimensionality.) Consider a d-dimensional, zero-mean, spherical multivariate Gaussian distribution:
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}_d).$$
   Equivalently, each entry of $\mathbf{x}$ is drawn iid from a univariate standard normal distribution.

   In familiar small dimensions ($d \leq 3$), "most" of the vectors drawn from a multivariate Gaussian distribution will lie near the mean. For example, the famous 68–95–99.7 rule for $d = 1$ indicates that large deviations from the mean are unusual. Here we will consider the behavior in larger dimensions.

   - Draw 10 000 samples from $p(\mathbf{x})$ for each dimension in $d \in \{1, 5, 10, 50, 100\}$, and compute the length of each vector drawn: $y_d = \sqrt{\mathbf{x}^\top \mathbf{x}} = (\sum_i^d x_i^2)^{1/2}$. Estimate the distribution of each y_d using either a histogram or a kernel density estimate (in MATLAB, `hist` and `ksdensity`, respectively). Plot your estimates. (Please do not hand in your raw samples!) Summarize the behavior of this distribution as d increases.
   - The true distribution of y_d^2 is a chi-square distribution with d degrees of freedom (the distribution of y_d itself is the less-commonly seen chi distribution). Use this fact to compute the probability that $y_d < 5$ for each of the dimensions in the last part.
   - For $d = 1\,000$, compute the 5th and 95th percentiles of y_d. Is the mean $\mathbf{x} = \mathbf{0}$ a representative summary of the distribution in high dimensions? This behavior has been called "the curse of dimensionality."

2. (Bayesian linear regression.) Consider the following data:
$$\mathbf{x} = [-2.26, -1.31, -0.43, 0.32, 0.34, 0.54, 0.86, 1.83, 2.77, 3.58]^\top;$$
$$\mathbf{y} = [1.03, 0.70, -0.68, -1.36, -1.74, -1.01, 0.24, 1.55, 1.68, 1.53]^\top.$$
   Fix the noise variance at $\sigma^2 = 0.5^2$.

   - Perform Bayesian linear regression for these data using the polynomial basis functions $\phi_k(x) = [1, x, x^2, \dots x^k]^\top$ for $k \in \{1, 2, 3\}$, in each case using the parameter prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$. Evaluate and plot the posterior means $\mathbb{E}[\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2]$ on the interval $x_* \in [-4, 4]$ for each model. Also plot the posterior mean plus-or-minus two times the posterior standard deviation:
$$\mathbb{E}[\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2] \pm 2\sqrt{\operatorname{var}[\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2]}.$$
     This is a pointwise 95% credible interval for the regression function. Where is the pointwise uncertainty the largest?
   - Compute the marginal likelihood of the data for each of the basis expansions above: $p(\mathbf{y} \mid \mathbf{X}, k, \sigma^2)$. Which model explains the data the best?

3. (Optimal design for Bayesian linear regression.) Consider the data from the last problem, and suppose we have selected the quadratic model corresponding to $k = 2$ (do not assume that this is the answer to the last part of the last question). Imagine we are allowed to evaluate

the function at a point x' of our choosing, giving a new dataset $\mathcal{D}' = \mathcal{D} \cup \big\{(x', y')\big\}$ and a new posterior for the parameters $p(\mathbf{w} \mid \mathcal{D}', \sigma^2) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_{\mathbf{w}\mid\mathcal{D}'}, \boldsymbol{\Sigma}_{\mathbf{w}\mid\mathcal{D}'})$. We hope to select the location x' to best improve our current model, under some quality measure.

Assume that we ultimately wish to predict the function at a grid of points

$$\mathbf{x}_* = [-4, -3.5, -3, \dots, 3.5, 4]^\top.$$

We select the squared loss for a set of predictions $\hat{\mathbf{y}}_*$ at these points:

$$\ell(\mathbf{y}_*, \hat{\mathbf{y}}_*) = \sum_i \big((y_*)_i - (\hat{y}_*)_i\big)^2;$$

therefore, we will predict using the new posterior mean $\hat{\mathbf{y}}_* = \mathbf{X}_* \boldsymbol{\mu}_{\mathbf{w}\mid\mathcal{D}'}$.

- Given a potential observation location x', derive a closed-form expression for the expected loss $\mathbb{E}\big[\ell(\mathbf{y}_*, \hat{\mathbf{y}}_*) \mid x', \mathcal{D}\big]$. Note: this does not require integration over y'! (What is the expected squared deviation from the mean?)
- Plot the expected loss over the interval $x' \in [-4, 4]$. Where is the optimal location to sample the function?

Note: this approach of actively selecting where to sample a function to maximize some utility function is known as *active learning* in machine learning and *optimal experimental design* in statistics. Bayesian decision theory provides a convenient and consistent framework for performing active learning with a variety of objectives.

4. (Woodbury matrix identity.) The *Woodbury matrix identity* is a very useful result. Let $\mathbf{A}$ be an $(n \times n)$ matrix, let $\mathbf{U}$ and $\mathbf{V}$ be $(n \times k)$ matrices, and let $\mathbf{C}$ be a $(k \times k)$ matrix. Then:

$$(\mathbf{A} + \mathbf{U}\mathbf{C}\mathbf{V}^\top)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C} + \mathbf{V}^\top\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top\mathbf{A}^{-1}.$$

This result is useful when you already have the inverse of a matrix $\mathbf{A}$ and want to know the inverse after a rank-k adjustment. When $k \ll n$, the Woodbry matrix identity can be considerably faster than direct inversion!

- Prove this result.
- Use this result to rewrite the posterior covariance of the weight vector $\mathbf{w}$ in Bayesian linear regression (as written in the notes to lecture 5) in a simpler form.

5. (Laplace approximation.) Find a Laplace approximation to the Gamma distribution:

$$p(\theta \mid \alpha, \beta) = \frac{1}{Z}\theta^{\alpha-1}\exp(-\beta\theta).$$

Plot the approximation against the true density for $(\alpha, \beta) = (2, 1/2)$.

The true value of the normalizing constant is

$$Z = \frac{\Gamma(\alpha)}{\beta^\alpha}.$$

If we fix $\beta = 1$, then $Z = \Gamma(\alpha)$, so we may use the Laplace approximation to estimate the Gamma function. Analyze the quality of this approximation as a function of α.