Assignment #2

Due: 23:59pm October 4, 2013

Problem 1 (10pts, Murphy)

Given that we have an estimate $\hat{\boldsymbol{w}}$ of the weights of a linear regression model with Gaussian noise, show that the MLE of the error variance is given by

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\mathsf{T} \hat{\boldsymbol{w}})^2.$$

Problem 2 (15pts)

One intuitive way to summarize a probability density is via the mode, as this is the "most likely" value in some sense. A common example of this is using the maximum *a posteriori* (MAP) estimate of a model's parameters. In high dimensions, however, the mode becomes less and less representative of typical samples. Consider variates from a D-dimensional zero mean spherical Gaussian with unit variance:

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_D, \mathbb{I}_D),$$

where $\mathbf{0}_D$ indicates a column vector of D zeros and $\mathbb{I}_D$ is a $D \times D$ identity matrix.

1. Compute the distribution that this implies over the distance of these points from the origin. That is, compute the distribution over $\sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$, if $\boldsymbol{x}$ is a realization from $\mathcal{N}(\mathbf{0}_D, \mathbb{I}_D)$. (Hint: Consider transformations of a Gamma distribution.)

2. Make a plot that shows this probability density function for several different values of D, up to $D = 100$.

3. Make a plot of the cumulative distribution function (CDF) over this distance distribution for $D = 100$. A closed-form solution may be difficult to compute, so you can do this numerically. (Hint: In Matlab, look up `cumtrapz`.)

4. From examining the CDF we can think about where most of the mass lives as a function of radius. For example, most of the mass for $D = 100$ is within a thin spherical shell. From eyeballing the plot, what are the inner and outer radii for the shell that contains 90% of the mass in this case?

Problem 3 (15pts)

Consider a mixture model for a one-dimensional random variable X arising from the following generative procedure:

- With probability $\frac{1}{2}$, X is Gaussian with zero mean and variance four.
- With probability $\frac{3}{8}$, X is Laplace-distributed with location five and scale two.
- With probability $\frac{1}{8}$, X is uniform on $(-2, -1.5)$.

1. Write the PDF for this mixture model.
2. Produce a plot of the probability density.
3. Draw 500 samples from this distribution and produce a normalized histogram.
4. Produce a plot that shows the 95% central credible region. This may require numeric integration.
5. Produce a plot that shows the 95% high posterior density region. This may require discretization and/or optimization.

Problem 4 (30pts)

Here are some simple data to regress:

```
x = [-1.87 -1.76 -1.67 -1.22 -0.07 0.11 0.67 1.60 2.22 2.51]'
y = [0.06 1.67 0.54 -1.45 -0.18 -0.67 0.92 2.95 5.13 5.18]'
```

Construct a Bayesian linear regression model using a basis of your choosing (e.g., polynomial, sinusoids, radial basis functions). Choose priors that seem sensible for the regression weights and the Gaussian noise.

1. Identify your basis and your priors and explain why you chose them.
2. Plot the data, as well as several typical posterior samples of the function given the data.
3. Plot the 95% central credible interval region of the predictive density as a function of x. That is, produce a plot that shows the "tube" containing most of the functions that are consistent with the data under your model.
4. There are probably different numbers of basis functions you could choose for your model. For example, you could choose the order of polynomial, or how many radial basis functions to put down. Fix a choice of the noise, and then produce a bar plot for several different such hypotheses that shows their marginal likelihoods. Do the data support one hypothesis over the others? Which one?

Problem 5 (30pts, Hastie et al., Murphy)

In this problem, we'll apply logistic regression to a data set of spam email. These data consist of 4601 email messages, from which 57 features have been extracted. 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, e.g., "business", "free", etc.
- 6 features in $[0, 100]$, giving the percentage of characters in the email that match characters on a list containing, e.g., "$", "#", etc.
- 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.

There are files `spam.train.dat` and `spam.test.dat` (available on the course website) in which each row is an email. There are 3000 training and 1601 test examples. The final column in each file indicates whether the email was spam.

1. Apply ℓ_2-regularized logistic regression. Use cross-validation to determine an appropriate regularization penalty. Report your procedure and the value you find. What training and test performance do you get with this value?
2. There are different ways one might preprocess the data. One typical thing to do is to "standardize" each input feature so that it has mean zero and variance one. Do this standardization and evaluate the model again. How do your results change?
3. In some data, what matters most is whether the data are zero or non-zero, and not what the actual value is. Transform the features to be binary in this way and retrain the model as above.
4. Alternatively, some features are best represented via their logs. Transform the features, retrain the model and report results as above.