COS 424: Interacting with Data

Homework #4 — Spring 2007
Regression — Due: Wednesday, April 18

Written Exercises

See the course website for important information about collaboration and late policies, as well as where and when to turn in assignments. Be sure to show your work and justify your answers.

Problem E-1

Consider a mixture of Gaussians model defined by K means $\mu_1, \ldots, \mu_K$, variance σ^2, and proportions $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_K \rangle$. In such a model, each (real-valued) X_n is generated as follows: First, one of the mixture components $Z_n \in \{1, \ldots, K\}$ is chosen at random according to $\boldsymbol{\pi}$ (so that $Z_n = z$ with probability π_z). Then, given that $Z_n = z$, X_n is chosen according to a Gaussian distribution with mean μ_z and variance σ^2. Note that only X_n is visible; Z_n is hidden. We assume that $\sigma > 0$ is known and fixed.

a. Give a graphical model depiction of this process, including the parameters.

b. Given data $X_1, \ldots, X_N$, describe in detail the EM algorithm for estimating $\mu_1, \ldots, \mu_K$ and $\boldsymbol{\pi}$.

c. Argue that as $\sigma^2 \to 0$, this algorithm approaches the K-means algorithm.

d. Argue directly that as $\sigma^2 \to 0$, the EM objective approaches the K-means objective.

Problem E-2

As is usual for linear regression, suppose we are given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ where $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^n$ (with components x_{ij}). In this problem, we seek linear models of the form $\hat{f}(\mathbf{x}) = w_0 + \mathbf{w} \cdot \mathbf{x}$ where w_0 is the scalar intercept term, and $\mathbf{w} = \langle w_1, \ldots, w_n \rangle$ is a (column) vector of weights over the n inputs. Consider the problem in ridge regression of minimizing

$$\sum_{i=1}^{m} (w_0 + \mathbf{w} \cdot \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2 . \tag{1}$$

Here, as in Hastie et al. (but unlike in class), we include an explicit intercept term w_0, but omit this term from the regression penalty.

a. Suppose *for this part only* that $\sum_{i=1}^{m} x_{ij} = 0$ for all j. Let $\mathbf{X}$ be the $m \times n$ matrix of all inputs in which the i-th row is equal to (the transpose of) $\mathbf{x}_i$, and let $\mathbf{y}$ be the (column) vector whose i-th entry is y_i. Show that the solution of (1) is given by

$$\begin{aligned} \hat{w}_0 &= \frac{1}{m} \sum_{i=1}^{m} y_i \\ \hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

b. Returning to the general case (in which the input vectors do not sum to zero), let

$$a_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$$

and define $\mathbf{x}'_i$ by $x'_{ij} = x_{ij} - a_j$. Note that, after centering in this fashion, the new input vectors sum to zero so that the technique in the last part can be applied. Show that minimizing (1) is equivalent to minimizing

$$\sum_{i=1}^{m}(w'_0 + \mathbf{w}' \cdot \mathbf{x}'_i - y_i)^2 + \lambda \left\|\mathbf{w}'\right\|_2^2. \tag{2}$$

In other words, if $\hat{w}_0, \hat{\mathbf{w}}$ is the solution that minimizes (1), and $\hat{w}'_0, \hat{\mathbf{w}}'$ is the solution that minimizes (2), show that $\hat{w}_0 + \hat{\mathbf{w}} \cdot \mathbf{x} = \hat{w}'_0 + \hat{\mathbf{w}}' \cdot \mathbf{x}'$ for any $\mathbf{x}$ and its transform $\mathbf{x}'$. Moreover, given a solution $\hat{w}'_0, \hat{\mathbf{w}}'$ to (2), show explicitly how to transform it directly into a solution $\hat{w}_0, \hat{\mathbf{w}}$ to (1).

c. Suppose that the inputs are both centered *and* scaled. In other words, suppose we instead define $\mathbf{x}'_i$ by $x'_{ij} = (x_{ij} - a_j)/s_j$ for some constants s_j. Show that the minimization problems (1) and (2) need no longer be equivalent (in the sense described above). Show nevertheless how a solution $\hat{w}'_0, \hat{\mathbf{w}}'$ to (2) can be transformed back into $\hat{w}_0, \hat{\mathbf{w}}$, not necessarily a solution to (1), but for which $\hat{w}_0 + \hat{\mathbf{w}} \cdot \mathbf{x} = \hat{w}'_0 + \hat{\mathbf{w}}' \cdot \mathbf{x}'$ for any $\mathbf{x}$ and its transform $\mathbf{x}'$.

Problem E-3

In this problem, we will show that ridge regression can be kernelized by showing that the minimization criterion can be rewritten in terms only of inner products of input vectors. Assume a training set as in the last problem.

Let $\mathbf{w} \in \mathbb{R}^n$ be any vector, and let $\alpha_1, \ldots, \alpha_m$ be chosen to minimize

$$F(\alpha_1, \ldots, \alpha_m) = \left\|\mathbf{w} - \sum_{i=1}^{m} \alpha_i \mathbf{x}_i\right\|_2^2.$$

Let

$$\mathbf{v} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i.$$

a. Prove that $\mathbf{v} \cdot \mathbf{x}_i = \mathbf{w} \cdot \mathbf{x}_i$ for all i.
(Hint: consider the partial derivatives of F.)

b. Prove that $\|\mathbf{v}\|_2 \leq \|\mathbf{w}\|_2$ with equality if and only if $\mathbf{v} = \mathbf{w}$.

c. Consider minimizing

$$\sum_{i=1}^{m}(\mathbf{w} \cdot \mathbf{x}_i - y_i)^2 + \lambda \left\|\mathbf{w}\right\|_2^2 \tag{3}$$

where $\lambda \geq 0$. Show that this minimization problem must always have a solution $\mathbf{w}$ which is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_m$. (You do not need to show that *every* solution has this form, just that there always exists at least one solution in this form.)

d. Substituting $\sum_i \alpha_i \mathbf{x}_i$ for $\mathbf{w}$, rewrite (3) so that the input vectors $\mathbf{x}_i$ only appear as inner products with other input vectors, thus showing that this minimization problem can be kernelized.

Problem E-4

In class, we showed that if X and Y are random variables with Y real-valued, then

$$\mathrm{E}\left[(\hat{f}(X) - Y)^2\right]$$

is minimized over all choices of the function $\hat{f}$ when $\hat{f}(x) = \mathrm{E}\left[Y|x\right]$, that is, when $\hat{f}(x)$ is the expected value of Y for a given x. In this problem, we will see what happens if we instead use

$$\mathrm{E}\left[|\hat{f}(X) - Y|\right].$$

In particular, we will see that this expectation of the "absolute loss" is minimized over all choices of $\hat{f}$ when $\hat{f}(x)$ is equal to the *median* of Y given x. (We say that m is a median of a real-valued random variable Z if $\Pr[Z \geq m] \geq 1/2$ and $\Pr[Z \leq m] \geq 1/2$. Note that the median is not always unique.)

To simplify the problem, we fix x (as we did in class) and also assume that Y is concentrated on a finite set of values. Thus, the problem can be reformulated as follows: Let $c_1 < c_2 < \cdots < c_\ell$ be the finite set of values in Y's range, and let Y be equal to c_j with probability p_j (where we assume without loss of generality that the p_j's are all strictly positive).

Prove that

$$\mathrm{E}\left[|\hat{f} - Y|\right]$$

is minimized over $\hat{f} \in \mathbb{R}$ if and only if $\hat{f}$ is equal to a median of Y.
(Hint: First determine the value $\hat{f}$ that minimizes $\mathrm{E}\left[|\hat{f} - Y|\right]$ when $\hat{f}$ is restricted to lie in the confined range $[c_i, c_{i+1}]$. Then use your answer to prove the general result.)

Experiments and programming in R

Data and code: The data, helper code and function templates for the problems below can be obtained by following the links on the webpage for this assignment. Once downloaded into a suitable location, just type `source("helper.R")` to load all of the datasets described below (other than `co2` which is already built in to R). In every case, this will load training and test sets as appropriately named data frames. The quantity to be predicted will always be the last column of these data frames.

What to turn in: You should turn in the file `stubs.R` with all of the functions filled in. Also turn in any other files that you may have created and used to complete this assignment, for instance containing additional R functions; be sure to include the R code that you used to complete each of the problems. These should all be submitted electronically using moodle, as should your predictions for optional Problem R-5, if you choose to complete it. Your written answers to the various problems, including any plots that you generated, should be submitted in hard copy together with your written exercises.

Problem R-1

Implement the ridge regression algorithm as described in Section 3.4.3 of Hastie et al. (as well as Problem E-2 above). Use exactly the method they describe in which an implicit intercept term is always included in the regression, but is always omitted from the penalty term used by ridge regression; use the "centering" technique they describe for this purpose. Also, each input (a.k.a feature or dimension or variable) should be scaled to have unit variance on the training set. In other words, if the training examples are $\mathbf{x}_1, \ldots, \mathbf{x}_m$ where x_{ij} is the j-th input value of $\mathbf{x}_i$, then each x_{ij} should be replaced by $(x_{ij} - a_j)/s_j$ where

$$a_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}$$

and

$$s_j = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_{ij} - a_j)^2}.$$

To make predictions on test examples, you will need to scale and center them in the same fashion (using the same constants a_j and s_j computed only on the training set).

Call your procedure `ridge()`; the parameters and return values for this procedure are detailed in the provided file called `stubs.R`.

Test your procedure to be sure it is working. For instance, you can try it on small data matrices for which the answer can be computed by hand. Write a short paragraph explaining what steps you took to be sure your code is working properly (this is crucial since you will be using this function throughout the rest of the assignment).

Problem R-2

The `co2` time series dataset records atmospheric concentrations of carbon dioxide monthly from 1959 to 1997. This dataset should already be built in to standard distributions of R

as the time-series object `co2`. (Be careful not to confuse it with the `CO2` dataset, also built in to R.) The goal is to estimate future CO_2 concentrations as a function of the year t. We will experiment with a number of bases, or sets of functions. For instance, the basis $\{t, t^2\}$ means that we are seeking an approximation of CO_2 concentration of the form

$$w_0 + w_1 t + w_2 t^2,$$

where t is the date in years, possibly with a fractional part (as returned, for instance, by `time(co2)`). (On this assignment, we always implicitly include the intercept term.)

Write a function called `run.co2()` that runs your ridge regression code on data collected through a specified cut-off year for a variety of bases. Details of what is required for this procedure are given in `stubs.R`.

a. Run your ridge regression procedure on each basis given below with the regression parameter λ set equal to 10^{-6}:

   - $\{t\}$
   - $\{t, t^2\}$
   - $\{t, t^2, \ldots, t^{20}\}$
   - $\{t, t^2, \ldots, t^{50}\}$
   - $\{t, t^2, \cos(2\pi t)\}$
   - $\{t, t^2, \sin(2\pi t)\}$
   - $\{t, t^2, \sin(2\pi t), \cos(2\pi t)\}$

   Use all data from the start of the recording period in 1959 through the end of 1974 as training data; use the data collected from the beginning of 1975 through the end of the recording period in 1997 as a test set. For each basis below, record (and turn in) the root mean squared error (RMSE) on the test set, where the RMSE of a model $\hat{f}$ on a set $(x_1, y_1), \ldots, (x_m, y_m)$ is given by

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{f}(x_i) - y_i)^2}.$$

   The RMSE on a test set is considered a reasonable measure of the accuracy of the predictions.

   Also make a plot showing actual CO_2 concentrations as a function of the year, superimposed with the concentrations (both training and test) predicted by the model that you built for that basis. (You might wish to use colors to distinguish predictions in the training set from predictions in the test set.)

b. Experiment with other choices of the parameter λ and/or other bases. Turn in the results of these experiments.

c. Briefly discuss the results, including the following questions: Which model gives the best predictions? How do the results make sense and fit your expections (and how do they not)? If it were the beginning of 1975 and you only had data through that date, which model would have seemed the "most reasonable" to have chosen for predicting future CO_2 concentrarions, and how accurate would that model actually have been? How does working with this kind of time-series data differ from i.i.d. data?

Problem R-3

The *housing* dataset was constructed from the 1990 US census. Here, the goal is to predict the median price of a house in a given region based on demographic composition and the state of the housing market in that region. Detailed descriptions of the variables used in this and the other datasets are available on the main webpage for this assignment. After following the instructions above, this dataset will be loaded into the variables `house.train` and `house.test`.

The *abalone* dataset is concerned with predicting the age of abalone (a shellfish) from physical measurements. The age of abalone can be determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope — a tedious and time-consuming task. Here, we instead attempt to predict the age using other measurements, which are easier to obtain, such as sex, length, diameter, etc. After following the instructions above, this dataset will be loaded into the variables `abalone.train` and `abalone.test`.

a. On the housing dataset, run ridge regression on the provided training and testing sets for a variety of values of the regularization parameter λ. Explore this parameter using exponentially varying values such as $\ldots, 4, 2, 1, 0.5, 0.25, \ldots$. Plot test RMSE as a function of $\log_{10}(\lambda)$. Your goal, of course, is to minimize the test RMSE, and you may need to do some exploring to find the "interesting" range of λ where this quantity is smallest. (It is not necessary that your `ridge()` function work for very small values of λ due to numerical limitations of R when solving linear equations. This may limit the range of values you can search.)

b. Repeat the experiment in the last part for the abalone dataset.

c. Briefly discuss the results above, including the following questions: How sensitive is the performance on the test set to the choice of λ for each of the datasets? In other words, if we are slightly off in our choice of λ, how much will this affect test performance? How do the plots you made for housing and abalone differ qualitatively, or how are they similar? Speculate on possible explanations for these differences or similarities, and what these mean in practice when choosing λ.

d. Using the housing test set, make a scatterplot comparing predicted median housing prices to actual housing prices. In other words, make a scatterplot with one point for each example in the test set where the x-coordinate is the actual median housing price, and the y-coordinate is the price predicted by your `ridge()` procedure when run on the housing dataset with the best choice of λ (as determined above). Briefly discuss the quality of the fit as reflected on this scatterplot.

e. Manually explore the weight coefficients returned by your `ridge()` procedure when run on the housing dataset with the best choice of λ (as determined above). Discuss what can or cannot be inferred from these coefficients regarding the factors that do or do not influence the median price of houses in a region.

Problem R-4

Implement k-fold cross validation as a means for estimating test error using only the training set. Your procedure should be called `cross.val()`, and should call `ridge()` k times in the standard manner to obtain k estimates of the test error which are averaged and returned by the function. Details of what is required are given in the `stubs.R` file.

a. Run 10-fold cross validation on the housing training set for the same set of λ values you used in part (a) of the last problem. On a single plot, show the test RMSE as a function of $\log_{10}(\lambda)$ as estimated using cross validation. On the same plot, also show your results from part (a) of the last problem so that the estimated and actual test error numbers can be compared.

b. Repeat the experiment in the last part for the abalone dataset, comparing your results to those obtained in part (b) of Problem R-3.

c. Discuss these plots. How good a job is cross validation doing as a method for estimating test error? How good a job is cross validation doing as a method for selecting the best choice of λ?

Problem R-5 *(optional and for extra credit)*

The *computer-activity* dataset records various performance measures from a Sun Sparcstation, such as the bytes read or written from system memory. The goal is to predict the percentage of time that the cpu is running in user mode. After following the instructions above, this dataset will be loaded into the variables `comp.train` and `comp.test`. Note that we are *not* providing target values for the test set (all of these have been set to zero). For this part of the assignment, we are asking you to use whatever techniques you wish to come up with the most accurate predictions you can for the test set. After the deadline, we will compare your submitted results to the actual test values, and we will post the results on-line.

For this part of the homework, you can use some of the techniques explored on this problem set. Or you can try different methods such as feature selection, lasso, nearest-neighbors, decision trees, kernel regression, and more. You can use any of the functions built in to R, provided you understand what they are doing. You are welcome and encouraged to try out your own ideas. (But please do not use any methods that go against the spirit of this assignment, such as searching for this dataset on the web.)

You should save and submit a file with your name (or a pseudonym), a one-sentence description of the technique used, and a list of predictions on test examples. A function called `save.prediction.file()` has been provided for this purpose (see `helper.R` for details on how to use it). The saved file should be submitted via moodle. We will use the information you provide to generate a public, on-line compilation of the results. If you wish to remain anonymous, you do not need to provide your real name in the file you generate, but can instead use a pseudonym of your choice.

Also, write and turn in a brief description of the approach you followed to generate your predictions, including plots you might have used along the way (for instance, for choosing λ if using ridge regression), as well as an explanation of why you tried what you did. Finally, be sure to turn in any code that you wrote for this part.