Machine Learning for Finance – Problem Set 5

Neal Parikh

April 10, 2018

Instructions. Do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you discussed any problems with other students, indicate that in your solutions.

1. *Cross validation and feature selection.* Fed up with friends who insist you "aren't curing cancer" with your statistical algorithms, you decide to turn your attention to applying machine learning methods to cancer research.

   You obtain a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where the input vector $x_i \in \mathbf{R}^n$ contains the *expression levels* of a large number of genes in a tumor tissue and the class label $y_i \in \{0, 1\}$ indicates whether the tumor is cancerous or benign. (The x_i are called *gene expression data* or *microarray data* in computational biology, and obtaining them is called *gene expression profiling.*) The goal is to fit a binary classifier to diagnose cancerous tumors using gene expression levels. Generally, the number of tissue samples N is relatively small, *e.g.*, $N = 50$, while the number of genes n is in the thousands, *e.g.*, $n = 5000$.

   You decide to begin with a simple approach called *nearest centroid classification.* Here, we compute the elementwise averages $\overline{x}^{\text{benign}}$ and $\overline{x}^{\text{cancer}}$ of input vectors in classes 0 and 1, respectively; we then classify a new query point x^{new} based on which of these it is closer to.

   In order to make the procedure more efficient, you decide to preprocess the data with a feature selection procedure. You find the 100 features (genes) with the highest correlation with the class labels and throw away the measurements on the remaining 4900 genes, *i.e.*, each data point $(\tilde{x}_i, y_i)$ now has $\tilde{x}_i \in \mathbf{R}^{100}$ rather than $\mathbf{R}^{5000}$.

   You now want to estimate the test set performance of the classifier. You split your simplified dataset $\tilde{\mathcal{D}} = \{(\tilde{x}_1, y_1), \ldots, (\tilde{x}_N, y_N)\}$ into $K = 5$ folds, then compute the cross-validation error of the nearest centroid classifier.

   Is this a correct use of cross validation? Why or why not?

2. *Reverse linear regression.* Suppose you have a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ with both $x_i, y_i \in \mathbf{R}$. There are a number of statistics that measure the linear relationship between x and y, such as *Pearson's correlation coefficient*, which is the covariance between x and y divided by the product of their standard deviations. It ranges from -1 (perfect negative linear correlation) to 1 (perfect positive linear correlation), with 0 being no linear correlation.

   Clearly, given the definition above, the correlation coefficient between x and y is the same as that between y and x. This is one case where it does not matter which of x and y are treated as the 'input' or the 'output'.

Is it equivalent to regress x on y (*i.e.*, treat x as inputs and y as outputs) and to regress y on x (*i.e.*, treat y as inputs and x as outputs)? What is the difference between these two approaches, if any?

Hint. It may help to draw a picture.

3. *Interpreting model fitting results.* Five different models are fit using the same training data set, and tested on the same (separate) test set, which has the same size as the training set. The RMS (square root of MSE) prediction errors for each model, on the training and test sets, are reported below. Comment briefly on the results for each model. You might mention whether the model's predictions are good or bad, whether it is likely to generalize to unseen data, or whether it is overfit. You are also welcome to say that you don't believe the results, or think the reported numbers are fishy.

Model	Train RMS	Test RMS
A	1.355	1.423
B	9.760	9.165
C	5.033	0.889
D	0.211	5.072
E	0.633	0.633

4. *Debugging learning algorithms.*

   (a) Suppose you train a regularized logistic regression classifier for handwritten digit classification by solving

   $$\text{maximize} \quad \textstyle\sum_{i=1}^N \log p(y_i \mid x_i; w) - \lambda \|w\|_2^2,$$

   with variable $w \in \mathbf{R}^n$. You measure the classification error rate of your model on both your training set and a holdout cross-validation set. Suppose that your model achieves low training error but high test set error. How should you adjust λ (*i.e.*, increase or decrease) in order to improve the model, and why?

   (b) Suppose that on the same handwritten digit classification task, you decide to switch to a soft-margin support vector machine, which involves solving

   $$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0, \end{array}$$

   with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, $t \in \mathbf{R}^N$.

   Using the same features as the previous logistic regression model, you find that the SVM gives both high training error and test error. How should you adjust λ to improve the model, and why?

   (c) Consider fitting a ridge regression model, *i.e.*, carrying out MAP estimation for a linear regression model with the parameter prior $w \sim \mathrm{N}(0, \tau^2 I)$. If the training error is much lower than the test error, should you increase or decrease τ to try to improve test error?

(d) Consider a classification problem, and define the training error to be the fraction of training examples misclassified by logistic regression. We generally expect a supervised learning algorithm to do better as the number of training examples N increases. Is it true or false that we expect the training error to decrease as N increases? Explain.

5. *Prediction contests.* Several companies have run prediction contests open to the public. Netflix ran the best known contest, offering a $1M prize for the first prediction of user movie rating that beat their existing method's RMS prediction error by 10% on a test set. The contests generally work as follows (although there are several more complex variations on this format). The company posts a public data set, that includes the regressors or features and the outcome for a large number of examples. They also post the features, but not the outcomes, for a (typically smaller) test data set. The contestants, usually teams with obscure names, submit predictions for the outcomes in the test set. Usually there is a limit on how many times, or how frequently, each team can submit a prediction on the test set. The company computes the RMS test set prediction error (say) for each submission. The teams' prediction performance is shown on a *leaderboard*, which lists the 100 or so best predictions in order.

   Discuss such contests in terms of model validation. How should a team check a set of predictions before submitting it? What would happen if there were no limits on the number of predictions each team can submit?

6. *Ridge regression.* The ridge regression problem is to solve

$$\text{minimize} \quad \|Ax - b\|_2^2 + \lambda \|x\|_2^2,$$

with variable $x \in \mathbf{R}^n$, where $A \in \mathbf{R}^{m \times n}$ and $\lambda > 0$.

(a) *The normal equations.* The linear least squares problem has the closed form solution

$$x^\star = (A^T A)^{-1} A^T b \tag{1}$$

when the columns of the feature matrix A are linearly independent. Derive a similar closed form expression for the solution of the ridge regression problem.

(b) *High-dimensional estimation.* If $m > n$ (*i.e.*, there are more features than training examples), we cannot use the estimator (1) because $A^T A$ is singular. One major benefit of regularization methods like ridge regression and the lasso is that they work in this regime, which is common in modern applications (*e.g.*, when working with gene expression data). Explain why we can still use the ridge estimator even when $m > n$.

(c) *Bayesian interpretation.* Show that MAP estimation in a linear regression model with a $\mathrm{N}(0, \tau^2 I)$ prior on the parameters involves solving a ridge regression problem. Show that the ridge estimator is also the posterior mean.

Note. Even though MAP estimation is not a fully Bayesian approach, we still say that interpreting a problem as carrying out MAP estimation under a particular prior amounts to providing a 'Bayesian interpretation' of the original problem. This is common usage, but it is important to understand what is meant.

7. *MAP estimates and weight decay.* Consider using a logistic regression model, with model weights $w \in \mathbf{R}^n$, for a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. Let $\hat{w}^{\mathrm{ML}}$ and $\hat{w}^{\mathrm{MAP}}$ be the maximum likelihood and maximum a posteriori estimates of w, respectively, where the MAP estimate is obtained assuming a $\mathrm{N}(0, \tau^2 I)$ prior on w.

   Prove that $\|\hat{w}^{\mathrm{MAP}}\|_2 \leq \|\hat{w}^{\mathrm{ML}}\|_2$. This property is the reason why the use of this type of regularization is sometimes referred to as *weight decay.*

8. *ℓ_1- and ℓ_2-norm approximation by a constant vector.* What is the solution of the norm approximation problem with one scalar variable $x \in \mathbf{R}$,

$$\text{minimize} \quad \|x\mathbf{1} - b\|,$$

   for the ℓ_1- and ℓ_2-norms?

9. *Least absolute deviations.* Just as we can use the ℓ_1 penalty $\|x\|_1$ instead of Tikhonov regularization $\|x\|_2^2$, we can consider the use of the loss function $\|Ax - b\|_1$ rather than the squared loss $\|Ax - b\|_2^2$. The resulting (unregularized) model is known as *least absolute deviations*, and it provides a criterion different from least squares for fitting a linear regression line. (It can also be given a probabilistic interpretation as carrying out maximum likelihood estimation in a particular model.)

   Based on your understanding about the difference between ℓ_1 and quadratic penalties, briefly explain how you might expect least absolute deviations to differ from standard least squares.

10. *Dirichlet-multinomial model.* One of the main questions in Bayesian modeling is how to choose a prior distribution appropriate in a given situation. One of the most convenient approaches is to choose a prior distribution that is *conjugate* to a given likelihood. The density of the conjugate prior follows the same general functional form as the likelihood, and has the property that the posterior lies in the same family as the prior, just with adjusted parameters. Though the conjugate prior may not be the best choice for a given problem, its use greatly simplifies many of the calculations that appear in Bayesian statistics. In this problem, we will carry out these derivations in detail for one example of interest.

    Consider modeling coin flips. The natural choice of likelihood is the Bernoulli(p) distribution, where the distribution parameter p is the probability of flipping heads and so must lie in the interval $[0, 1]$. The *beta distribution* is a distribution on $[0, 1]$, and draws from this distribution are probabilities that can be used as the parameter of a Bernoulli distribution. The beta distribution has the density

$$p(x) \propto x^{\alpha-1}(1-x)^{\beta-1},$$

    with parameters α, $\beta > 0$. The normalization constant is a complicated function and is referred to as the *beta function* $\mathrm{B}(\alpha, \beta)$; it can be expressed in terms of *gamma functions* as

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

    where the gamma function Γ is a continuous extension of the factorial function to all real numbers. The function satisfies the condition $\Gamma(n) = (n-1)!$ for any positive integer n and the recurrence $\Gamma(x+1) = x\Gamma(x)$ for any positive real x. (There is an integral representation for Γ, but this is not relevant to the problem, and it is not expanded further because there are numerical methods available to evaluate it in standard scientific computing environments.)

This situation can be generalized to the multinomial distribution over K outcomes; in this case, the conjugate prior is the *Dirichlet distribution*, which is a continuous distribution over the probability simplex, *i.e.*, vectors in $\mathbf{R}^K$ that are nonnegative and sum to 1. Samples from the Dirichlet distribution amount to weights on a K-sided die. The Dirichlet density is

$$p(x) \propto \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

with parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$. In this case, its normalization constant is called the *multivariate beta function* $\mathrm{B}(\alpha)$ and is given by

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}.$$

Explicitly, then, the Dirichlet density is

$$p(x) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}.$$

We now come to the problem. Consider the generative model

$$\begin{aligned} \varphi &\sim \text{Dirichlet}(\alpha) \\ z \,|\, \varphi &\sim \text{Multinomial}(\varphi), \end{aligned}$$

with $z \in [K]$, and suppose we have a dataset $\mathcal{D} = \{z_1, \ldots, z_N\}$ generated i.i.d. from the model above. Explicitly, the likelihood is given by

$$p(\mathcal{D} \,|\, \varphi) = \prod_{k=1}^{K} \varphi_k^{N_k},$$

where N_k is the number of observations in $\mathcal{D}$ with class k. (Note that the N_k are precisely the sufficient statistics of the multinomial distribution.)

(a) *Posterior distribution.* Show that

$$p(\varphi \,|\, \mathcal{D}) = \text{Dirichlet}(\alpha_1 + N_1, \ldots, \alpha_K + N_K).$$

(b) *Predictive distribution.* Show that the posterior predictive distribution is given by

$$p(y = k \,|\, \mathcal{D}) = \frac{N_k + \alpha_k}{N + \mathbf{1}^T \alpha}.$$

(c) Give an English interpretation of the two expressions above.

(d) Give a Bayesian interpretation of Laplace smoothing.

Hint. There are some complicated integrals that appear in calculations involving these distributions; it is useful to express them in terms of gamma functions. A *beta integral* is an integral in the form

$$\iint_\Delta x^p y^q \, dx \, dy,$$

where the domain of integration Δ is the probability simplex. It can be expressed as

$$\iint_\Delta x^p y^q \, dx \, dy = \frac{p!q!}{(p+q+2)!} = \frac{\mathrm{B}(p+1, q+1)}{p+q+2},$$

where $\mathrm{B}(u, v)$ is the beta function.

Similarly, an integral in the form

$$\int \cdots \int_\Delta \prod_{i=1}^n x_i^{\alpha_i - 1} \, dx_1 \cdots dx_n$$

is called a *Dirichlet integral of type 1.* It can be expressed as

$$\int \cdots \int_\Delta \prod_{i=1}^n x_i^{\alpha_i - 1} \, dx_1 \cdots dx_n = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)}.$$

11. *Laplace smoothing.* Suppose you have N observations $z_1, \ldots, z_N$ of a Bernoulli(p) random variable. Let

$$\hat{p} = \frac{1}{N} \sum_{i=1}^N z_i$$

be the maximum likelihood estimate of p, and let $\hat{p}'$ be the Laplace smoothed estimate of p. Is $\hat{p}'$ closer to $1/2$ than $\hat{p}$?