Homework 2: Spam, Spam, Spam, Decision Theory, and Spam

36-465/665, Spring 2022

Due at 6 pm on Thursday, 3 February 2022

Agenda: First exposure to classification, and classification errors; more practice with the difference between in-sample error and generalization error; first exposure to decision theory (for classification).

The first few problems are all applied, and are about classifying e-mails as spam or not. The last few problems are theory, easing you in to thinking about prediction, and in particular classification, as a decision problem. The two halves complement each other, but they can be done in either order.

In Q2 through Q5, we are using the `spam` data set in the `ElemStatLearn` library. Download the package from CRAN, and load the data set into memory with `data(spam)`. The goal is to learn to classify e-mail as either spam or real mail. There are 58 columns: the first 57 are features (see `help(spam)`, and [https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names]), and the last one is a categorical variable ("factor" in R-speak), named `spam`, with two values, `email` and `spam`. We'll treat `spam` as the "positive" class, and `email` as the "negative" class. Each of the 4601 rows is a different e-mail.

In Q6 and Q7, we are trying to predict a binary variable Y, which for definiteness we can say is either 0 or 1. Our prediction will be $\hat{Y}$, also 0 or 1. The prediction is based on an input variable ("feature") X, and $\Pr(Y = 1|X = x) = p(x)$ is the probability that Y is 1 when $X = x$.

1. **Online questions** (10) are on Canvas and due at 6 pm on Monday.
2. **Base rates** To see whether a classifier is actually working, we should compare it to a constant classifier which always predicts the same class, no matter what the input features actually are.
   a. (1) What fraction of the e-mails are actually spam?
   b. (3) What should the constant classifier predict, in order to be as accurate as possible?
   c. (2) What is the error rate of the constant classifier?
3. **Training and testing** (4) Divide the data set at random into a training set of 2301 rows and a testing set of 2300 rows. Check that the two halves do not overlap, and that each has the right number of rows. (Do *not* print lists of row numbers.) What fraction of each half is spam? Pick three of the other variables and check that they have *roughly* the same distribution in both sets. Show your code for all of this, including comments explaining your approach and choices. *Hint*: see solutions to Q8a in HW 1.
4. **Our first logistic regression** Fit a logistic regression model, with the outcome being the `spam` variable, on all 57 of the features. (Do not report the coefficients unless you have some specific point to make about them.) Use only the training data in the fitting.
   a. (2) For each point in the training data, use the logistic regression to calculate the probability that that point is spam, say p_i for data point i. Include a histogram of the predicted probabilities. Comment on the shape of the histogram.
   b. (3) For each point in the training data, classify the point as `spam` if the predicted probability of being spam is ≥ 0.5, and not spam otherwise. (In symbols: $\hat{Y}_i = \mathbb{I}\{p_i \geq 0.5\}$.) What is the error rate? Is this better than the constant, baseline classifier of Q2? Is that surprising?
   c. (2) Take the sub-set of training points which are actually spam. What fraction of them did you classify as not spam in Q4b? (This is the false negative rate, $\Pr\left(\hat{Y} = 0|Y = 1\right)$.)
   d. (2) Take the sub-set of training points which are actually not spam. What fraction of them did you classify as spam in Q4b? (This is the false positive rate, $\Pr\left(\hat{Y} = 1|Y = 0\right)$.)

e. (2) What fraction of data points that you classified as spam in Q4b are actually spam? (This is the positive predictive value, $\Pr\left(Y = 1 | \hat{Y} = 1\right)$.)

f. (2) Find the average negative log probability of the actual labels, $-\frac{1}{n}\sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log (1 - p_i)$. How is this related to the log likelihood of the model? *R hint*: the `ifelse()` command can help here.

5. **Testing our first logistic regression** Keep *exactly* the model you fitted to the training set in Q4, but now apply it *only* to the points in the testing set. Do *not* re-calculate any parameters, only predictions. In Q5b–Q5f, say whether the model is doing better on the training set, better on the testing set, or equally well on both. *R hint*: Use the `predict()` function, with the option `type="response"` (why?).
   a. (3) Repeat Q4a. How similar are the two histograms?
   b. (3) Repeat Q4b.
   c. (3) Repeat Q4c.
   d. (3) Repeat Q4d.
   e. (3) Repeat Q4e.
   f. (3) Repeat Q4f.
6. **Binary classification with 0-1 loss** In this problem, we incur a loss of 0 if $Y = \hat{Y}$, and a loss of 1 if $Y \neq \hat{Y}$. Say that $\hat{Y} = 1$ when $p(x) \geq 0.5$ and $\hat{Y} = 0$ when $p(x) < 0.5$.
   a. (4) Show that the probability of mis-classifying is $p(x)$ if $p(x) < 0.5$, and $1 - p(x)$ if $p(x) > 0.5$.
   b. (3) Show that the probability of mis-classifying can be written as $\min(p(x), 1 - p(x))$ for all x.
   c. (2) Explain why we can set $\hat{Y} = 0$ or $\hat{Y} = 1$ when $p(x) = 0.5$ without changing the probability of error.
   d. (3) Show that the probability of mis-classifiying can be written as $\frac{1}{2} - \left|p(x) - \frac{1}{2}\right|$ for all x.
   e. (5) Find an expression for the risk of the optimal classifier.
   f. (3) When will the risk be zero? Explain.
7. **Decision boundaries for binary classification** In this problem, assume that we are trying to predict a binary variable Y, which for definiteness can be either 0 or 1. We have a 2×2 matrix L_{ij} which tells us the loss we incur when we predict j but the reality is i.
   a. (3) Find an expression for the expected loss of setting $\hat{Y} = 0$, conditional on $X = x$. Your answer should be a function of $p(x)$ and the elements of the L matrix.
   b. (3) Find an expression for the conditional expected loss of setting $\hat{Y} = 1$, similarly to Q7a.
   c. (3) Give a criterion for when it is better to set $\hat{Y} = 0$ and when it is better to set $\hat{Y} = 1$. You should be able to manipulate your answer to be of the form "Predict $\hat{Y} = 1$ when $p(x) >$" some expression involving the elements of the L matrix.
   d. (3) Explain, in words, what your answer in Q7c says about when you should be indifferent between saying $\hat{Y} = 0$ and $\hat{Y} = 1$.
   e. (3) Show that when $L_{00} = L_{11} = 0$ and $L_{01} = L_{10} > 0$, the boundary from Q7d is precisely at $p(x) = 1/2$. Explain this in words.
8. **Connecting theory and practice** (3) Why was it reasonable to use a threshold of 0.5 in Q4 and Q5? When would it make sense to require passing a much higher threshold before classifying something as spam? *Hint*: Q6 and Q7.
9. **Timing** (1) How long, roughly, did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All plots and tables are generated by code included in the R Markdown file. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for.

Extra credit 1: Randomized decisions We're in the same setting as Q7, but now, instead of always ("deterministically") setting $\hat{Y} = 0$ or $\hat{Y} = 1$ depending on which side of the boundary x is on, we *randomly* set $\hat{Y} = 1$ with probability $q(x)$, which is not necessarily equal to the probability $p(x)$ that $Y = 1$.

a. (1) Find an expression for the expected loss at a given x. It should involve $q(x)$, $p(x)$, and the elements of the L matrix.

b. (2) Hold x fixed, and find an expression for the value of $q(x)$ that will minimize the expected loss. Your answer should involve $p(x)$ and the elements of the loss matrix.

c. (2) Does the possibility of a randomized rule lead to any improvements over what we can do with deterministic rules? That is, are there any times where we'd get better performance with a randomized rule than a deterministic one?

Extra credit 2: But what if we're wrong? (5) In Q5b, you calculated the classification accuracy (on the test set) applying a threshold of 0.5 to the predicted probabilities. Re-do the classification where the threshold is allowed to vary, in symbols $\hat{Y}_i = \mathbb{I}\{p_i \geq t\}$, for $t \in [0, 1]$. Plot the accuracy versus t. Is it actually maximized when $t = 0.5$? If not, can you explain why it might not be?