Homework 4: Linear Classifiers

36-462/662, Spring 2020

Due at 10 pm on Thursday, 13 February 2020

AGENDA: Gaining familiarity with linear classifiers.

Our problems this week are a mixture of some (basic) theory about classifiers and linear classifiers, and some applications of linear methods to the COMPAS data set from last week. On the course website, you will find a file giving row numbers for a random 80% of the data set. Use those rows as your training set, and the other 20% of the rows as your testing set, rather than making your own split.

There are no online reading questions this week.

1. *Minimum-error classification* In this problem, you will prove that the way to minimize the probability of mis-classification is to always predict the most probable class.

   Let Y be the class, which is binary (0 or 1), and $\vec{X}$ the vector of features. Let $\mathbb{P}\left(Y = 1|\vec{X} = \vec{x}\right) = p(\vec{x})$. Consider a classifier which makes a *randomized* prediction, predicting 1 with probability $q(\vec{x})$. Further, suppose that the actual class and that the prediction are conditionally independent given $\vec{X}$.

   (a) (4) For each fixed $\vec{x}$, show that the conditional probability of mis-classification, $r(\vec{x})$, is $q(\vec{x}) + p(\vec{x}) - 2p(\vec{x})q(\vec{x})$.

   (b) (3) Plot the error rate r as a function of q in the interval $[0, 1]$, for $p = 0.1$, $p = 0.3$, $p = 0.5$, $p = 0.6$ and $p = 0.9$. (Your plot should have five different lines or curves, one for each value of p, clearly distinguished on the plot.) Based on your plot, which values of q minimize the error rate r for each these four values of p?

   (c) (2) Show that, holding p fixed, the derivative of r with respect to q is never zero (unless $p = 1/2$), and has the same sign for all values of q. How does this relate to the plot you just made?

   (d) (3) Using your previous work in this problem, show that if $p(\vec{x}) > 1/2$, then $r(\vec{x})$ is minimized when $q(\vec{x}) = 1$, and that if $p(\vec{x}) < 1/2$, then $r(\vec{x})$ is minimized when $q(\vec{x}) = 0$. Is $q = 1/2$ optimal when $p = 1/2$?

2. *The prototype method is a linear classifier* Recall that the prototype method classifies a vector $\vec{x}_0$ by assigning it to the class with the closest "prototype" vector. With two classes, the prototype for the positive class $\vec{c}_+$ is the average of the positive training vectors, $\vec{c}_+ = \frac{1}{n_+}\sum_{i:\ y_i=1} \vec{x}_i$, and likewise for the negative class, $\vec{c}_- = \frac{1}{n_-}\sum_{i:\ y_i=0} \vec{x}_i$, where n_+ and n_- are the number of positive and negative training vectors.

   (a) (5) Show that the prototype method is really a linear classifier of the form $\mathbf{1}\left\{b + \vec{x}_0 \cdot \vec{w} \geq 0\right\}$, and find b and $\vec{w}$ in terms of $\vec{c}_+$ and $\vec{c}_-$.

   Hint: Try writing "$\vec{x}$ is closer to $\vec{c}_+$ than to $\vec{c}_-$" as an inequality between two distances, and then squaring both sides.

   (b) (5) Find the dual weights for the prototype method. That is, expressing the classifier as $\mathbf{1}\left\{b + \sum_{i=1}^{n} \alpha_i \vec{x}_i \cdot \vec{x}_0 \geq 0\right\}$, find the α_i.

   (c) (2) How many support vectors are there? That is, how many of the training vectors $\vec{x}_i$ have non-zero weights α_i?

3. *Logistic regression's classification boundary* In this problem, $p(\vec{x}) = \mathbb{P}\left(Y = 1 | \vec{X} = \vec{x}\right)$.

   (a) (5) Suppose we fit a logistic regression with two features, so our model is $\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. We want to find the boundary where $p(\vec{x}) = 1/2$, so that the log odds are equal to 0. What is an equation for x_2 in terms of x_1 on the boundary?

   (b) (3) Now suppose there are d features. Show that, on the boundary,

$$x_d = -\frac{1}{\beta_d}\left(\beta_0 + \beta_1 x_1 + \ldots \beta_{d-1} x_{d-1}\right)$$

4. Load the COMPAS data.

   (a) (5) Using the training data, create a plot of priors against age. Use shape or color to indicate whether the arrestee was a recidivist.

   Hint: Because many arrestees have the same combination of age and priors, you may want to use the `jitter` function to add a small amount of noise to the plot, so that different people will show up as distinct points. (See the examples at the end of `help(jitter)`.)

   (b) (5) Based on your plot, if you had to use a linear classifier with these features, where, *roughly*, would you put the boundary? Would you predict "recidivist" above or below the line? Explain.

   Hint: You can use `abline()` to add a straight line to the plot.

   Request: Try to do this one "by eye", rather than adjusting your line after doing later problems.

5. *Prototype method in practice*

   (a) (3) Using the training data, and just the two features of age and priors, what are the vectors for the average recidivist and the average non-recidivist?

(b) (3) Add the prototypes to your plot from problem 4a. Going by this plot, do you think you will have a higher false positive rate or false negative rate?

(c) (3) Classify every data point in the testing set according to which prototype the point is closer to. What is the confusion matrix? What is the error rate? What are the false positive and false negative rates?

   Hint: Use `knn()`, from the `FNN` library. What should the `train` and `cl` arguments be? What should k be?

6. *Linear probability models*

   (a) (3) On the training data, fit a linear regression for recidivism, using the two features of age and priors. What are the coefficients and their standard errors?

   (b) (3) What do your coefficients mean?

   (c) (2) What do the standard errors of your coefficients mean?

   (d) (3) What proportion of the points in the training set have estimated probabilities of recidivism < 0? What proportion have estimated probabilities > 1? What are these proportions in the testing set?

   (e) (3) Calculate predictions for every point in the testing set, and threshold the probabilities at 0.5. (Ignore the fact that some of the probabilities are impossible.) What is the confusion matrix? What is the error rate? What are the false positive and false negative rates?

7. *Logistic regression*

   (a) (3) Using the training data, fit a logistic regression for recidivism, using the two features of age and priors. What are the coefficients and their standard errors?

   Hints: Look at the slides for 4 February for a little worked example of fitting a logistic regression to two features. Make sure you're fitting a logistic regression, and not a linear regression!

   (b) (3) What do your coefficients mean?

   (c) (2) What do the standard errors on your coefficients mean?

   (d) (3) Re-create your plot from problem 4a, and add the classification boundary you'd get from your estimated logistic regression. How well does it match up to the line you "drew by eye" in problem 4b?

   Hint: Problem 3a.

   (e) (5) For each data point in the test set, use your logistic regression to predict the probability of recidivism. Classify points as recidivists if their probability is ≥ 0.5. What is the confusion matrix? What is the error rate? What are the false positive and false negative rates?

Hint: When you fit a model with `glm()`, the `predict()` function takes a `type` argument. The default value is `"link"`, which gives predictions on the scale of the "link function", which for logistic regression is the log odds. On the other hand, `predict(my.glm, newdata, type="response")` will give predictions on the scale of the response variable, which for logistic regression will be probabilities. You can use either here, just don't mix them up.

(f) (5) Use a range of probability thresholds between 0 and 1, and calculate the false negative rate and false positive rate you would get at each threshold (on the testing set). Plot the false negative rate against the false positive rate. Is there a trade-off between false negatives and false positives?

(g) (4) How many people in the testing set have predicted probability of recidivism in the range $[0, 0.1)$? (That is, ≥ 0 and < 0.1?) What fraction of people in this "bin" are actually recidivists? Should this fraction be between 0 and 0.1? Repeat for the $[0.1, 0.2)$ bin, and so on. (You may find that some bins are empty, that is there are no predictions in that range; skip them.) Plot the actual fraction of recidivists in each bin against the average predicted probability in that bin. For full credit, make the area of each plotted point proportional to the number of people in the bin. Why should this be a straight, diagonal line if the model is right? What does the departure from that shape tell you about how the model is imperfect?

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. All plots and tables are automatically generated by code included in the source file, rather than being made by hand in an external program, or hard-coded. 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 quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all questions are answered with coherent text, and never with raw computer code or its output.

Extra credit (5): Fit a support vector machine with a Gaussian kernel to predict recidivism using age and priors. Explain how you picked the scale (bandwidth) of the Gaussian kernel. Make a plot showing the prediction of the SVM over the age/priors space, with support vectors distinguished from other data points. Report the confusion matrix, the error rate, and the false positive and false negative rates.