Homework 11

36-465/665, Spring 2021

Due at 6 pm on Thursday, 22 April 2021

Agenda: Working with kernel machines

1. *Data preparation.* Accompanying the COMPAS data file (on the class website) is a separate file giving the row numbers to be used in the training set (80%), the rest being reserved for a testing set.

    a. (5) Create a new column in the data frame which is $+1$ for those who are arrested for violence within 2 years, and -1 for those who are not. Use the old column to check that this has been done correctly. *All further references to "recidivism" in this homework are to the new column.*

    b. (5) Generate summary statistics for (i) age, (ii) number of priors and (iii) recidivism status between the training and the testing sets. How closely do they match? How closely should they match?

    c. (5) Calculate the correlation between (i) age and priors, (ii) age and recidivism, and (iii) priors and recidivism, in both the training and and the testing sets. How closely do the correlations match? How closely should they match?

    d. (5) What was the point of Q1b and Q1c?

2. *Kernel ridge regression, take 1* Use kernel ridge regression, with Y = recidivism, and X = (age, priors). Use a Gaussian (radial basis function) kernel, with a bandwidth of 1. Fit the kernel machines to the training data for a range of λ (penalty) values between 0.01 and 100. (Use at least 5 values of λ, but really the more points you can plot in the curves below better, provided the cost in computer time stays reasonable.)

    a. (5) For each value of λ, use the results of HW10 Q1 to calculate the empirical Rademacher complexity. (Strictly speaking, it's an upper bound on the empirical Rademacher complexity.) Plot the Rademacher complexity as a function of λ, and comment on the shape of the curve.

    b. (3) For each value of λ, calculate the mean squared error of the machine's predictions on the training set. Plot the error curve, i.e., the MSE as a function of λ, and comment on the shape of the curve.

    c. (5) Repeat Q2b, but calculating the mean squared error on the testing set. (Be careful to not re-fit the machines to the testing set!) Does this error curve resemble the sum of the curves from Q2a and Q2b? Should it?

    d. (5) What value of λ, among the ones you tried out, would you recommend using, with this loss function?

    e. (5) Give at least one reason why the squared error loss function is bad for this application.

3. *Kernel ridge regression, take 2* If we're predicting a variable y which is ± 1, with a real-valued prediction m, the **margin loss** $\ell(y, m) = -ym$.

    a. (4) Explain the saying "the margin loss is negative at correctly-classified points and positive at incorrectly-classified points."

    b. (4) Explain the saying "the margin loss likes to put the boundary between classes far from every data point".

c. (3) Repeat Q2b but displaying the average margin loss for the different versions of ridge regression. Comment on the resulting curve.

d. (5) Repeat Q2c but for the margin loss. Comment.

e. (5) Would you chose a different value of λ than in Q2d?

f. (5) Give at least one reason why using the hinge loss makes more sense in classification problems than does squared error. Give at least one reason why building a kernel machine using ridge regression, but evaluating it using hinge loss, is less than ideal. (You don't have to suggest a better alternative; we'll look at some next week.)

4. (5) Explain how (if at all) you could use the data to select a good bandwidth in the kernel. (You don't need to implement your idea.)

5. *Ethics* It would be bad teaching to have you work with a controversial data set like this, and not ask you to think about some of the reasons why there's controversy. Remember, the main use of models like the COMPAS model is to help judges decide which people who have been accused, but not convicted, of crimes can be safely released before their trial.

   a. (5) Is it fair, or just, to use *age* as an attribute in this kind of prediction? Give (at least) one reason for *and* one reason against. You may want to consider both fairness to person arrested, and justice towards the members of the various larger communities or groups affected by this decision. (Even if you strongly believe that it is, or is not, right to include age in such a model, you should be able to come up with at least one *good* reason for the opposite position.)

   b. (5) Is it fair, or just, to use *number of prior convictions* in a model like this? Again, give at least one reason for and one reason against.

   c. (5) Should statistical models like this have any role in making this kind of legal decision at all? Again, give at least one reason for and one reason against.

6. (1) How much time 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 quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Extra credit (10 points total): Visualizing basis functions

Lecture 20 gave some examples of displaying the (approximate) basis functions implicit in using a Gaussian kernel on this data, but it did that using one the one attribute of age. In this extra credit exercise, you'll do the same thing but using both age *and* priors, so the basis functions are two-dimensional and not one-dimensional.

General hint for this question: Look at the .Rmd file for Lecture 20.

a. Select a random subset of 200 records from the training set, and use it for the rest of this question. Why is using a subset, rather than the whole training set, helpful in this question? Would there be anything wrong or misleading in using the whole training set?
b. Using a Gaussian (radial basis function) kernel with a bandwidth of 1, on the two attributes of age and prior count together, find the kernel matrix $\mathbf{K}$ on your subset. Display the matrix as either a 2D image (using color to indicate values in the matrix), or as a 3D surface (using height). Comment.
c. Make a plot of the eigenvalues of $\mathbf{K}$. Comment. *Hint:* All of the eigenvalues should be ≥ 0 (why?), but your computer may calculate some as extremely small negative numbers (like -10^{-16}) due to rounding error; set those to zero.

d. Make a plot visualizing the *first* (largest-eigenvalue) eigenvector of $\mathbf{K}$, multiplied by the square root of its eigenvalue. (Why the square root?) This should be either a 2D or a 3D plot. If you chose a 2D plot, the x axis should be age, the y axis should be priors, and color should indicate the value of the eigenvector at that combination of age and priors. (Some of these might be misisng values.) If it is a 3D plot, the x axis should be age, the y axis should be priors, and the z axis should indicate the value of the eigenvector. Comment on the result. *R hint*: For a 3D plot, the easiest option may be the `scatterplot3d` function from the library of the same name.

e. Repeat (d) for the next three eigenvectors, multiplying each by the square root of its eigenvalue. Try to use the same color scheme and/or z range as in your plot in ECd, so that the visualizations are more directly comparable. Comment.

f. Repeat ECd–ECe for the four eigenvectors with the *smallest* non-zero eigenvalues. Comment.