Final Exam – Intro. to Computational Statistics

Unless otherwise specified, assume all α (p-value) thresholds to be 0.05, and all tests to be two-sided if that is an option. All calculations may be done with R or by hand unless otherwise specified. Please show and explain your work as much as possible, using latex for displaying all math.

Note that all problems are worth 3 points except problem 1, which is worth 7 points, and problems 7(a), 8(a), and 9(a), which are each worth 5 points.

Good luck!

1. You roll five six-sided dice. Write a script in R to calculate the probability of getting between 15 and 20 (inclusive) as the total amount of your roll (ie, the sum when you add up what is showing on all five dice). Exact solutions are preferable but approximate solutions are ok as long as they are precise.

2. Create a simulated dataset of 100 observations, where x is a random normal variable with mean 0 and standard deviation 1, and $y = 0.1 + 2 * x + \epsilon$, where epsilon is also a random normal error with mean 0 and sd 1. (One reminder: remember that in creating simulated data with, say, 100 observations, you need to use `rnorm(100)` for epsilon, not `rnorm(1)`, to ensure that each observation gets a different error.)

a. Perform a t test for whether the mean of Y equals the mean of X using R.

b. Now perform this test by hand using just the first 5 observations. Please write out all your steps in latex.

c. Using R, test whether the mean of Y is significantly different from 0.

d. Again using the first five obsevations, test by hand whether the mean of Y is different from 0.

e. Assuming the mean and sd of Y that you calculate from the first five observations would not change, what is the minimum *total* number of observations you would need to be able to conclude that the mean of Y is different from 0 at the p = 0.01 confidence level?

f. Verify (d) (approximately) by increasing the simulated data to the n you calculated in (e) that would be necessary. If the test of $Y = 0$ is still not significant, explain why. (Go back to using the original 100-observation dataset for g and h.)

g. Create a categorical (factor) variable c, where $c = 1$ if $x < -1$, $c = 3$ if $x > 1$, and $c = 2$ otherwise. Use R to perform an F test for whether the mean of y differs across these three groups.

h. Using the first three observations for each group, calculate the same F test by hand.

3. Generate a new 100-observation dataset as before, except now $y = 0.1 + 0.2 * x + \epsilon$

a. Regress y on x using R, and report the results.

b. Discuss the coefficient on x and its standard error, and present the 95% CI.

c. Use R to calculate the p-value on the coefficient on x from the t value for that coefficient. What does this p-value represent (be very precise in your language here)?

d. Discuss the F-statistic and its p-value, and calculate that p-value from the F statistic using R. What does this test and its p-value indicate?

e. Using the first five observations, calculate by hand the coefficient on x, its standard error, and the *adjusted* R^2. Be sure to show your work.

4. Now generate $y = 0.1 + 0.2 * x - 0.5 * x^2 + \epsilon$ with 100 observations.

a. Regress y on x and x^2 and report the results. If x or x^2 are not statistically significant, suggest why.

b. Based on the known coefficients that we used to create y, what is the effect on y of increasing x by 1 unit from 1 to 2?

c. Based on the coefficients estimated from 4(a), what is the effect on y of changing x from -0.5 to -0.7?

5. Now generate x_2 as a random normal variable with a mean of -1 and a sd of 1. Create a new dataset where $y = 0.1 + 0.2 * x - 0.5 * x * x_2 + \epsilon$.

a. Based on the known coefficients, what is the effect of increasing x_2 from 0 to 1 with x held at its mean?

b. Regress y on x, x_2, and their interaction. Based on the regression-estimated coefficients, what is the effect on y of shifting x from -0.5 to -0.7 with x_2 held at 1?

c. Regress the current y on x alone. Using the R^2 from this regression and the R^2 from 5(b), perform by hand an F test of the complete model (5b) against the reduced, bivariate model. What does this test tell you?

6. Generate a new variable y2 using the data from (5) which is 1 if $y > 0$ and 0 otherwise.

a. Perform a logistic regression of y2 on x, x_2, and their interaction, and interpret the results.

b. What is the effect of increasing x_2 from 0 to 1 with x held at its mean on the probability that y2 is 1?

7. Generate a dataset with 300 observations and three variables: f, x1, and x2. f should be a factor with three levels, where level 1 corresponds to observations 1-100, level 2 to 101-200, and level 3 to 201-300. Create x1 and x2 such that the first 100 observations have a mean of 1 for x1 and 1 for x2, each with a standard deviation of 2; the second 100 observations have a mean of 0 for x1 and 1 for x2, both with a standard deviation of 1; and the third 100 observations have a mean of 1 for x1 and 0 for x2, both with a standard deviation of 0.5.

a. Using the k-means algorithm, peform a cluster analysis of these data using a k of 3 (use only x1 and x2 in your calculations; use f only to verify your results). Comparing your clusters with f, how many datapoints are correctly classified into the correct cluster? How similar are the centroids from your analysis to the true centers?

b. Perform a factor analysis of this data using your preferred function. Using the scree plot, how many factors do you think you should include? Speculate about how these results relate to those you got with the cluster analysis.

8. Generate a dataset of 200 observations, this time with 90 independent variables, each of mean 0 and sd 1. Create y such that:

$$y = 2x_1 + ... + 2x_{30} - x_{31} - ... - x_{60} + 0 * x_{61} + ... + 0 * x_{90} + \epsilon$$

where ϵ is a random normal variable with mean 0 and sd 10. (Ie, the first 30 x's have a coefficient of 2; the next 30 have a coefficient of -1; and the last 30 have a coefficient of 0.)

a. Perform an elastic net regression of y on all the x variables using just the first 100 observations. Use 10-fold cross-validation to find the best value of λ and approximately the best value of α.

b. How accurate are your coefficients from (a)? Summarize your results any way you like, but please don't give us the raw coefficients from 90 variables.

c. Using the results from (b), predict y for the second 100 observations. How accurate is your prediction?

d. Attempt to compare the predictive accuracy here to the accuracy of a prediction made using regular multiple regression. Explain your results, including if the regular regression failed for any reason.

9. As in problem 6, use the data from 8 to generate a new y2 that is 1 if $y > 0$ and 0 otherwise.

a. Using the same process as in 8, estimate an SVM model of y2 on all the x variables for the first 100 variables. Use 10-fold cross-validation to select the best kernel.

b. Using the results from (a), predict y2 for the second 100 observations, and report your accuracy.