Statistical Learning Review
Exam Solutions

Student Name: _____
Professor: Nik Bear Brown

Rules:
1. NO COMPUTER, NO PHONE, NO DISCUSSION or SHARING.
2. Ask if you don't understand a question.
3. You may three 8½"×11" sheets of notes (you may use both sides, written or printed as small as you like).
4. Time allowed. Until end of class.
5. Bring pen/pencil. The midterm will be written on paper.

The exam will cover *"An Introduction to Statistical Learning with Applications in R"* chapters 1 to 5 as well as the additional regression and classification metrics that were covered in class.

Q1 (5 Points) Measuring the Quality of Fit

In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data. How is this done with regression, binary classification and multi-class classification? Give an example for regression, binary classification and multi-class classification? Give specific evaluation metrics.

Solution:

Any accurate example of regression, binary classification and multi-class classification metrics is fine.

A metric is a measure.

Regression metrics

MSE
Mean Squared Error. The "squared" bit means the bigger the error, the more it is punished. If your correct answers are 2,3,4 and your algorithm guesses 1,4,3, the absolute error on each one is exactly 1, so squared error is also 1, and the MSE is 1. But if your algorithm guesses 2,3,6, the errors are 0,0,2, the squared errors are 0,0,4, and the MSE is a higher 1.333.

deviance
Actually short for mean residual deviance. If the distribution is gaussian, then it is equal to MSE, and when not it usually gives a more useful estimate of error, which is why it is prefered to MSE.

RMSE
The square root of MSE. If your response variable units are dollars, the units of MSE is dollars-squared, but RMSE is back into dollars.

MAE

Mean Absolute Error. Following on from the MSE example, a guess of 1,4,3 has absolute errors of 1, so the MAE is 1. But 2,3,6 has absolute errors of 0,0,2 so the MAE is 0.667. As with RMSE, the units are the same as your response variable.

R2

R-squared, also written as R., and also known as the coefficient of determination.

RMSLE

The catchy abbreviation of Root Mean Squared Logarithmic Error. Prefer this to RMSE if an under-prediction is worse than an over-prediction.

Classification metrics

MAE, RMSE, etc are regression metrics. While they can be used for differences in probability they are not probabilistic distances.

misclassification

This is the overall error, the number shown in the bottom right of a confusion matrix. If it got 1 of 20 wrong in class A, 1 of 50 wrong in class B, and 2 of 30 wrong in class C, it got 4 wrong in total out of 100, so the misclassification is 4, or 4%.

mean_per_class_error

The right column in a confusion matrix has an error rate for each class. This is the average of them, so for the preceding example it is the mean of 1/20, 1/50, nd 2/30, which is 4.556%. If your classes are balanced (exactly the same size) it is identical to misclassification.

logloss

A probability for the answer being each category. The confidence assigned to the correct category is used to calculate logloss (and MSE). Logloss disproportionately punishes low numbers, which is another way of saying having high confidence in the wrong answer is a bad thing.

MSE

Mean Squared Error. The error is the distance from 1.0 of the probability it suggested. So assume we have three classes, A, B, and C, and your model guesses A with 0.91, B with 0.07, and C with 0.02. If the correct answer was A the error (before being squared) is 0.09, if it is B 0.93, and if C it is 0.98.

AUC

Area Under Curve.

See https://www.youtube.com/watch?v=OAl6eAyP-yo

Q2 (5 Points) Assume regression is being used to predict whether a student will graduate with honors or not. The dependent variable is **hon.** A yes as indicated by a 1 a no indicated by a 0. Assume the only independent variable is called math and is an integer representing a student's math score. The stats for the fit are shown in the table below.

hon	Coef.	Std. Err.	z
math	.1563404	.0256095	6.10
intercept	-9.793942	1.481745	-6.61

A. Write an equation that describes the model.
B. Is the coefficient math significant? How does one interpret the meaning of its value?
C. Is the coefficient intercept significant? How does one interpret the meaning of its value?

Solution:

A.

$\text{logit}(p) = \beta_0 + \beta_1 * \textbf{math}$

Any form of this equation is fine. (Like the one a little further down in this answer in terms or p rather than logit(p))

B. The coefficient math is significant as the z-score is 6.1 SD above the mean.

How do we interpret the coefficient for math? The coefficient and intercept estimates give us the following equation:

$\log(p/(1-p)) = \text{logit}(p) = -9.793942 + .1563404 * \textbf{math}$

The coefficient for **math** is the difference in the log odds. In other words, for a one-unit increase in the math score, the expected change in log odds is .1563404.

C. Is the coefficient intercept significant? How does one interpret the meaning of its value?

The coefficient intercept is significant as the z-score is 6.6 SD below the mean.

In this case, the estimated coefficient for the intercept is the log odds of a student with a math score of zero being in an honors class. In other words, the odds of being in an honors class when the math score is zero is exp(-9.793942) = .00005579. These odds are very low if we have a math score of 0 but it's unlikely that any student has a score that low. In turns out, all the test scores in the data set were

standardized around mean of 50 and standard deviation of 10. So the intercept in this model corresponds to the log odds of being in an honors class when **math** is at the *hypothetical* value of zero.

Q3 (5 Points) Calculate the increase in odds of receiving honors by going from a math score of 54 to a math score of 55 from the stats in Q2.

Solution:

Answer in terms of odds, log-odds or probability are OK for full credit.

Short Answer

The coefficient for **math** is the difference in the log odds. In other words, for a one-unit increase in the math score, the expected change in log odds is .1563404.

Going from a math score of 54 to a math score of 55 is just a one-unit increase in in the log odds (i.e. 0.1563404).

To go from log odds to odds just exponentiate it.

exp(.1563404) = 1.1692241.

For a one-unit increase in math score, we expect to see about 17% increase in the odds of being in an honors class. This 17% of increase does not depend on the value that math is held at, just a 1-unit increase.

Long but just fine Answer:

Or IF one wants to explicitly calculate the probability, and do more math, just plug the intercept beta zero and slope beta one in to the equation below; along with an X of 1000.

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

What is our estimated probability of `default` for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1 + e^{-10.6513+0.0055\times 1000}} = 0.006$$

P can be computed from the regression equation for a given value of X.

$$P = \frac{\exp(a+bX)}{1+\exp(a+bx)} = \frac{e^{a+bx}}{1+e^{a+bx}}$$

Note that the change (i.e. increase) will be the difference with an X=1 and X=0 (i.e. P(X=1) - P(X=0))

Please note that

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2} + 1} = \frac{1}{1 + b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

Either form of the equation is fine.

NOTE THIS IS NOT THE SAME AS THE EQUATION FOR MULTI-CLASS REGRESSION (i.e. Regression with more the two class) BELOW

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.

We can also show that the one unit increase does not depend on the value that math is held at.

Let's fix **math** at some value. We will use 54. Then the conditional logit of being in an honors class when the math score is held at 54 is

log(p/(1-p))(**math**=54) = – 9.793942 + .1563404 ***54**.

We can examine the effect of a one-unit increase in math score. When the math score is held at 55, the conditional logit of being in an honors class is

log(p/(1-p))(**math**=55) = – 9.793942 + .1563404***55**.

Taking the difference of the two equations, we have the following:

log(p/(1-p))(**math**=55) – log(p/(1-p))(**math** = 54) = .1563404.

We can say now that the coefficient for **math** is the difference in the log odds. In other words, for a one-unit increase in the math score, the expected change in log odds is .1563404.

Can we translate this change in log odds to the change in odds? Indeed, we can. Recall that logarithm converts multiplication and division to addition and subtraction. Its inverse, the exponentiation converts addition and subtraction back to multiplication and division. If we exponentiate both sides of our last equation, we have the following:

exp[log(p/(1-p))(**math**=55) – log(p/(1-p))(**math** = 54)] = exp(log(p/(1-p))(**math**=55)) / exp(log(p/(1-p))(**math** = 54)) = odds(**math**=55)/odds(**math**=54) = exp(.1563404) = 1.1692241.

So we can say for a one-unit increase in math score, we expect to see about 17% increase in the odds of being in an honors class. This 17% of increase does not depend on the value that math is held at.

Answer in terms of odds, log-odds or probability are OK for full credit.

Q4 (5 Points) Assume regression is being used to predict whether a student will graduate with honors or not. The dependent variable is **hon.** A yes as indicated by a 1 a no indicated by a 0. Assume that we have three independent variables: 1.) one called math and is an integer representing a student's math score. 2.) one called read and is an integer representing a student's reading score. And 3) a gender variable called female where the value (**female** = 1) means the gender is female. The stats for the fit are shown in the table below.

hon	Coef.	Std. Err.	z
math	.1229589	.0312756	3.93
female	.979948	.4216264	2.32
read	.0590632	.0265528	2.22
intercept	-11.77025	1.710679	-6.88

Write an equation that describes the model.

For all of the fitted parameters answer the following:

- Is the coefficient significant?
- How does one interpret the meaning of its value?

Solution:

This fitted model says that, holding **math** and **reading** at a fixed value, the odds of getting into an honors class for females (**female** = 1) over the odds of getting into an honors class for males (**female** = 0) is exp(.979948) = 2.66. In terms of percent change, we can say that the odds for females are 166% higher than the odds for males. The coefficient for **math** says that, holding **female** and **reading** at a fixed value,

we will see 13% increase in the odds of getting into an honors class for a one-unit increase in math score since exp(.1229589) = 1.13.

This fitted model says that, holding **math** and **reading** at a fixed value, the odds of getting into an honors class for females (**female** = 1) over the odds of getting into an honors class for males (**female** = 0) is exp(.979948) = 2.66 (0.979948 increase in the log-odds. In terms of percent change, we can say that the odds for females are 166% higher than the odds for males. The coefficient for **math** says that, holding **female** and **reading** at a fixed value, we will see 13% increase in the odds (0.1229589 increase in the log-odds) of getting into an honors class for a one-unit increase in math score since exp(0.1229589) = 1.13.

Q5 (5 Points) For the model in Q4 we and an interaction term of the two predictor variables female and math but removed the read predictor variable and fit the model. The stats for the fit are shown in the table below.

hon	Coef.	Std. Err.	z
female	-2.899863	3.094186	-0.94
math	.1293781	.0358834	3.61
femalexmath	.0669951	.05346	1.25
intercept	-8.745841	2.12913	-4.11

Write an equation that describes the model.

For all of the fitted parameters answer the following:

- Is the coefficient significant?
- How does one interpret the meaning of its value?

Solution:

logit(p) = log(p/(1-p))= β_0 + β_1***female** + β_2***math** + β_3***female*math**

In the presence of interaction term of **female** by **math**, we can no longer talk about the effect of **female**, holding all other variables at certain value, since it does not make sense to fix **math** and **femalexmath** at certain value and still allow female change from 0 to 1!

BUT full credit will be given if one discussed that the z-score is not significant.

In this simple example where we examine the interaction of a binary variable and a continuous variable, we can think that we actually have two equations: one for males and one for females. For males (**female**=0), the equation is simply

$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_2 * \textbf{math.}$

For females, the equation is

$\text{logit}(p) = \log(p/(1-p)) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) * \textbf{math.}$

Now we can map the logistic regression output to these two equations. So we can say that the coefficient for math is the effect of math when **female** = 0. More explicitly, we can say that for male students, a one-unit increase in math score yields a change in log odds of 0.13. On the other hand, for the female students, a one-unit increase in math score yields a change in log odds of (.13 + .067) = 0.197. In terms of odds ratios, we can say that for male students, the odds ratio is exp(.13) = 1.14 for a one-unit increase in math score and the odds ratio for female students is exp(.197) = 1.22 for a one-unit increase in math score. The ratio of these two odds ratios (female over male) turns out to be the exponentiated coefficient for the interaction term of **female** by **math**: 1.22/1.14 = exp(.067) = 1.07.

Q6 (5 Points) Think of some real-life applications for statistical learning.

(a) Describe a real-life application in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe a real-life application in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Solution:

Any reasonable examples are OK.

Q7 (5 Points) Describe the null hypotheses to which the p-values given in linear regression correspond.

Solution:

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

This question was from the practice exams.

Q8 (5 Points) What is the equation for multiple logistic regression?

Solution:

Multiple logistic regression when you have one nominal variable and two or more measurement variables. This is not the same as multi-class logistic regression. But based on the responses I feel that I didn't make this clear enough in class.

Because I feel the distinction was unclear in the lecture either the equation for multiple (multivariate) logistic regression or multi-class logistic regression was accepted. Please don't confuse the two but student were given credit either way.

Equation for multiple logistic regression

$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \ldots + \beta_n XN$

Equation for multi-class logistic regression (NOT multiple logistic regression)

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.

Q9 (5 Points) Describe backward step-wise selection.

Solution:

Backward Stepwise Selection

- Like forward stepwise selection, *backward stepwise selection* provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Directly from videos, slides and book.

Q10 (5 Points) Leave-*one*-out cross-validation (**LOOCV**) is a particular case of leave-*p*-out cross-validation with $p = 1$.

Explain how leave-*one*-out cross-validation (**LOOCV**) is implemented.

How would leave-*p*-out cross-validation with $p = 50$ differ from LOOCV? Any advantages or disadvantages?

Solution:

Explain how leave-*one*-out cross-validation (**LOOCV**) is implemented.

Any clear description of leaving out one test data point and using the rest of the data for training.

How would leave-*p*-out cross-validation with $p = 50$ differ from LOOCV?

Should mention it is faster.

Should mention that in leave-*one*-out cross-validation the n models are virtually the same training data. A little more variance in both the train and test data in leave-*p*-out cross-validation with $p = 50$.

Leave-one-out cross-validation

Leave-*one*-out cross-validation (**LOOCV**) is a particular case of leave-*p*-out cross-validation with $p = 1$.

The process looks similar to jackknife; however, with cross-validation one computes a statistic on the left-out sample(s), while with jackknifing one computes a statistic from the kept samples only.

Q11 (5 Points) Dr. Poindexter in his lecture on the logloss evaluation metric states "LogLoss heavily penalizes classifiers that are confident about an incorrect classification". Do you agree with Dr. Poindexter? Why or why not?

Solution:

It is true that "LogLoss heavily penalizes classifiers that are confident about an incorrect classification."

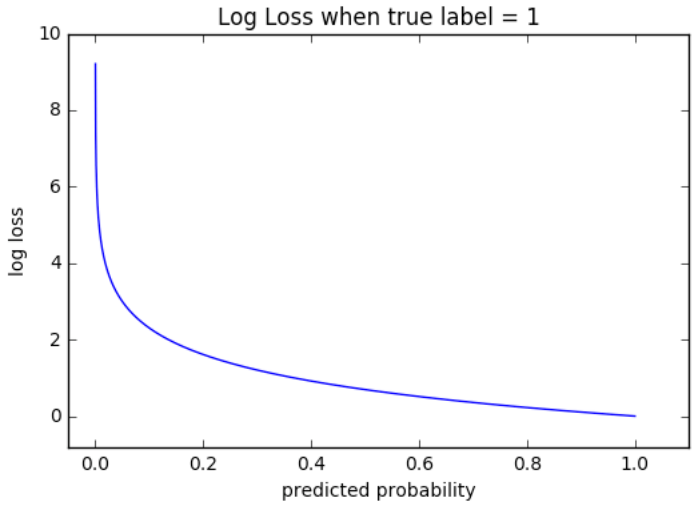
Yes it does.
3 points for getting that.

2 points for a good explanation of why or why not?

Classifiers return probabilities between 0 and 1.

Logarithmic loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1.

As the predicted probability approaches 1, log loss slowly decreases. As the predicted probability decreases, however, the log loss increases rapidly. Log loss penalizes both types of errors, but especially those predications that are confident and wrong! Lots of small errors may not add up to one large error.



Therefore, Log Loss heavily penalizes classifiers that are confident about an incorrect classification. For example, if for a particular observation, the classifier assigns a very small probability to the correct class then the corresponding contribution to the Log Loss will be very large indeed. Naturally this is going to have a significant impact on the overall Log Loss for the classifier. The bottom line is that it's better to be somewhat wrong than emphatically wrong when using this metric.

Q12 (5 Points) Assume one wants to use the three-class categorical variable high/medium/low as an independent variable in linear regression. How can we encode it?

Write an equation that describes the model. Assume we are fitting the intercept and one other continuous independent variable and our dependent variable is called 'y'.

Solution:

One would pick a base class and create dummy variables for the other two.

The multiple linear regression equation is as follows:

$y_i=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\epsilon_i$.

Where two of the slopes are dummy variables and the third is continuous independent variable and β_0 is the intercept.

Credit will be given for one-hot encoding but the multiple linear regression equation must be as follows:

$y_i=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\epsilon_i$.

Where three of the slopes are one-hot variables and the fourth is continuous independent variable and β_0 is the intercept.

Q13 (5 Points) I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I split the data into training and test sets. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta0 + \beta1X + \beta2 X^2 + \beta3 X^3 + e$.

Suppose that the true relationship between X and Y is linear. Consider the training and test residual sum of squares (RSS) for the linear regression, and also the training and test RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Solution:

This is question 3.7.4 part A from the chapter 3 exercises in the book, with a slight addition to the wording to provide a hint by adding a test set.

Without knowing more details about the training data, it is difficult to know which training RSS is lower between linear or cubic, as the true relationship between X *and* Y *is linear, we may expect the least squares line to be close to the true regression line, and consequently the RSS for the linear regression may be lower than for the cubic regression (or vice-versa) on the training data.*

On *test* data however, we would expect the overly complex (lower bias) model to overfit.

Q14 (5 Points) Suppose that we take a data set, divide it into training and test sets, and then try out two different classification procedures. First, we use model A and get an error rate of 20% on the training data and 25% on the test data. Then, we use model B and get an error rate of 22.5% on the training data and 22.5% on the test data.

Which model is preferred? Why?

Solution:

We care about the error rate on the test data, so we would prefer model B.

Q15 (5 Points) Let k be the number of estimated parameters in the model. Let n = the number of data points or equivalently, the sample size. Let L be the maximum value of the likelihood function for the model.

How do the Akaike information criterion (AIC) and Bayesian information criterion (BIC) differ?

Why would you use one or the other?

Solution:

Let k be the number of estimated parameters in the model. Let k be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

The BIC is formally defined as

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}).$$

3 Points for the equations

2 Points for discussion that BIC usually has a higher complexity penalty and ln(n) is almost always greater than k.

Q16 (5 Points) Is logistic regression a parametric or a non-parametric statistical learning approach. Why or why not?

Solution:

In parametric approaches one estimates parameter values, in non-parametric approaches one doesn't.

The equation for logistic below is a parametric equation. The slope and intercept are its parameters. That is we estimate the beta's (the parameters)

$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \ldots + \beta_n XN$

Q17 (5 Points) Write the equation for multiple linear regression with two continuous independent variables and one categorical independent variable. Are there any assumptions of the error model?

Solution:

One would pick a base class and create a dummy variable and the number of classes for the one categorical independent variable.

If one picks a two-class categorical independent variable, then there will be 2-1 or 1 dummy variable.

The multiple linear regression equation is as follows (for a 2-class categorical independent variable):

$y_i=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\epsilon_i$.

Where one of the slopes is the dummy variable and the two are the continuous independent variable and β_0 is the intercept.

Credit will be given for one-hot encoding but the multiple linear regression equation must be as follows (for a 2-class categorical independent variable):

$y_i=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\epsilon_i$.

Where two of the slopes are one-hot variables and the two are the continuous independent variable and β_0 is the intercept.

Q18 (5 Points) Suppose that we know about a population of students' weights that are normally distributed. Furthermore, suppose we know that the mean of the distribution is 170 pounds and the standard deviation is 20 pounds. Consider the following questions:

1. What is the *z*-score for 130 pounds?
2. What is the *z*-score for 300 pounds?
3. How many pounds corresponds to a *z*-score of 2.25?

Solution:

The formula that we will use is as follows: $z = (x - \mu)/ \sigma$

The description of each part of the formula is:

- x is the value of our variable
- μ is the value of our population mean.
- σ is the value of the population standard deviation.
- z is the z-score.

For the first question, we simply plug $x = 130$ into our z-score formula. The result is:

$(130 – 170)/20 = -2$

This means that 130 is two standard deviations below the mean.

The second question is similar. Simply plug $x = 300$ into our formula. The result for this is:

$(300 – 170)/20 = 6.5$

This means that 300 is 6.5 standard deviations above the mean.

For the last question, we now know our z -score. For this problem we plug $z = 2.25$ into the formula and use algebra to solve for x:

$2.25 = (x – 170)/20$

Multiply both sides by 20:

$45 = (x – 170)$

Add 170 to both sides:

$215 = x$

And we see that 215 pounds corresponds to a z-score of 2.25.

Q19 (5 Points) Dr. Poindexter in his lecture on over and under-fitting states that "resampling based measures such as cross-validation should never be preferred over mathematical measures such as Aikake's Information Criteria" when assessing model complexity. Do you agree with Dr. Poindexter? Why or why not?

Solution:

Do you agree with Dr. Poindexter? No.

Ultimately, we care about is how a model performs on out-of-sample data.

Resampling based measures such as cross-validation should be preferred over theoretical measures such as Aikake's Information Criteria.

AIC is ad hoc and what we care about is how a model performs on out-of-sample data; not a guesstimate complexity penalty.

Q20 (5 Points) Bootstrapping is a statistical method for estimating properties of an estimator (such as variance, confidence intervals) by measuring those properties when sampling from an approximating distribution.

Describe the algorithm in detail. Name another algorithm that uses bootstrapping.

Solution:

Bootstrapping

If your data has k rows, each bootstrap sample generated randomly samples the original *with replacement* the same size as the original dataset (k rows). We can generate as many bootstrap samples as we wish.

Any algorithm that uses bagging, such as a Random Forest, use bootstrapping. In fact, the name "bagging" stands for "bootstrap aggregation". Or just mentioning bagging or boosting is fine.

-2 if you don't mention *with replacement*.