CSYE 7245–
Big-Data Systems and Intelligence Analytics
Sample Quiz and Solutions One

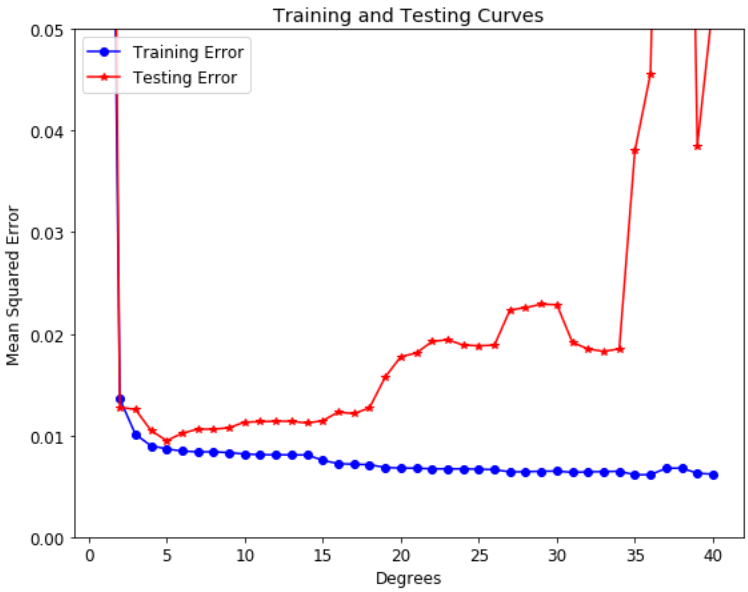
Q1 (15 Points) Each of n students gives their phone the Professor before a test. The Professor (being lazy) gives the phones back to the students in a random order. What is the expected number of students who get back their own phone?

Solution:

Phone-check problem

There are n phones and each person picks a phone uniformly at random hence each gets their right phone back with probability 1/n. Expectation is linear even when the random variables are dependent hence the mean of the total number of persons who get their right phone back is $n \times 1/n = 1$.

Q2 (15 Points) Polynomial models of various order (degree) were fit to the same data. Are any of the models below underfitting or overfitting? Name which models are under, overfit or correctly fit.



Solution:

Models well below degree 5 are clearly underfitting. The model of degree 5 is correctly fitting. Models well above degree 5 are clearly overfitting

Models 3 through 7 are close and are accepted as fitting as well.

Q3 (15 Points) Assume regression is being used to predict whether a student will graduate with honors or not. The dependent variable is **hon** which is yes as indicated by a 1 a no indicated by a 0. Assume the only independent variable called female and encoded with female as 1 and male as 0. The stats for the fit are shown in the table below.

hon	Coef.	Std. Err.	z
female	.5927822	.3414294	1.74
intercept	-1.470852	.2689555	-5.47

A. Write an equation that describes the model.
B. Is the coefficient female 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.
Writing it in an equation, the model describes the following linear relationship.

logit(p) = $\beta_0 + \beta_1$*female + error

y = $\beta_0 + \beta_1$*female is a -1. That is, missing proper form of the target.

or the correct alternate forms at the bottom of this solution are also OK

B.

A z-score of 1.74 corresponds to p-value of 0.083 so it is not significant at a 0.05 level.

The coefficient for **female** is the log of odds ratio between the female group and male group: log(1.809) = .593. So we can get the odds ratio by exponentiating the coefficient for female.

C.

A z-score of -5.47 corresponds to p-value of 0.000 so it is significant at a 0.05 level or 0.01 level or even well below that. In this case

The coefficient means log(p/(1-p)) = -1.471. What is p here? The target p is the overall probability of being in honors class (**hon** = 1) but the intercept of -1.471 is the log odds for males since male is the reference group (**female** = 0). Female = 0 is a male in this case.

To go from log odds to odds just exponentiate it.

exp(.-1.471) = 0.2296

Which is the odds of being in an honors class in male.

Note equations in this form for logistic regression are fine as well:

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.

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 female is held at.

Q5 (15 Points) Calculate the increase in the odds of receiving honors if one is female given the table in Q4.

Solution:

Note calculating the change (i.e. increase) in probability or just the probability or odds or log-odds are different.

Odds or log-odds answer.

The coefficient for **female** is the difference in the log odds. In other words, for a one-unit increase in the female score, the expected change in log odds is .5927822.

Going from a female score of 0 (i.e. male) to a female score of 1 is just a one-unit increase in in the log odds (i.e. 0.5927822).

To go from log odds to odds just exponentiate it.

exp(.5927822) = 1.809014

.

For a one-unit increase in female score, we expect to see about 81% increase in the odds of being in an honors class. That is being female, we expect to see about 81% increase in the odds of being in an honors class.

Probability answer.

Before trying to interpret the two parameters estimated above, let's take a look at the crosstab of the variable **hon** with **female** in the original data.

hon	female: male	female	Total
0	74	77	151
1	17	32	49
Total	91	109	200

We can manually calculate these odds from the table: for males, the odds of being in the honors class are (17/91)/(74/91) = 17/74 = .23; and for females, the odds of being in the honors class are (32/109)/(77/109) = 32/77 = .42. The ratio of the odds for female to the odds for male is (32/77)/(17/74) = (32*74)/(77*17) = 1.809. So the odds for males are 17 to 74, the odds for females are 32 to 77, and the odds for female are about 81% higher than the odds for males.

We can get this without knowing the original data, just the parameters of the fit.

Just plug the intercept beta zero and slope beta one in to the equation below; along with an X of

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.

Q6 (25 Points) Assume you have 3 binary classifiers (A,B,C) each with a 70% accuracy. You can view these classifiers for now as pseudo-random number generators which output a "1" 70% of the time and a "0" 30% of the time. The output of these classifiers is independent.

Create a machine learning ensemble algorithm that combines these 3 binary classifiers in to one prediction.

Describe your algorithm in detail.

What is accuracy of your algorithm?

Solution:

3 points for discussing some form of aggregation; like the average.

There are many ways of showing the ensemble of independent classifiers will increase the accuracy from 70% to around 78%.

For a majority vote with 3 members we can expect 4 outcomes:

We can also the formula for a discrete random variable based on a binomial distribution:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

Here 3 choose 2 with p=0.7 and 1-p=0.3 for getting 2 of 3.

And 3 choose 3 with p=0.7 and 1-p=0.3 for getting 3 of 3.

Or equivalently, use it to get P(1 of 3) and P(0 of 3)

Or we can calculate directly.

All three are correct P(3 of 3)

0.7 * 0.7 * 0.7

= 0.3429

Two are correct P(2 of 3)

0.7 * 0.7 * 0.3

+ 0.7 * 0.3 * 0.7

+ 0.3 * 0.7 * 0.7

= 0.4409

Two are wrong P(1 of 3)

0.3 * 0.3 * 0.7

+ 0.3 * 0.7 * 0.3

+ 0.7 * 0.3 * 0.3

= 0.189

All three are wrong P(0 of 3)

```
 0.3 * 0.3 * 0.3


= 0.027
```

We see that most of the times (~44%) the majority vote corrects an error. This majority vote ensemble will be correct an average of ~78% (0.3429 + 0.4409 = 0.7838).