**CSCE 623 Spring 2020: Machine Learning. In-Class Work, Day 6**

From Chapter 3: Linear Regression Part 1: “Evaluation Criteria Cheat Sheet”

There are many ways of determining which machine learning model parameters to use and measuring the performance of a model. One of the struggles with digesting chapter 3 is the presentation of many diagnostic and decision-making measures. Build a dictionary for each of these measures. Include the following for each measure:

* Math Formula, English Name/Definition & page number in book;
* Purpose / how to use it for decision-making / how to interpret;
* Benefits / assumptions / limitations / sensitivities

Example purposes: Use to estimate coefficients; Estimate accuracy of the model; hypothesis testing

Example benefits/assumptions/limitations: Expressed in the units of Y: Sensitive to the number of predictors

Residual sum of squares; pp 62. Use it to determine amount of error in a model. Sensitive to number of datapoints

MSE: Mean Squared Error; pp 29. Use it to determine amount of error in a model. NOT Sensitive to number of datapoints

TSS: Total Sum of Squared Errors; pp 70. Assuming no model, use the mean of the y’s to guess y. Sensitive to the number of datapoints, then compute the errors.

The standard error of the estimate of the mean; pp 65. Captures how far a single estimate of the mean will be, given sigma (the standard deviation of each of the realizations of y, assuming n observations are uncorrelated)

= {depends on which coefficient}

Standard error of the estimate of the jth coefficient; pp 68. Used to determine how close the coefficients are to their true values in the population phenomenon

Residual standard error; pp 66 for simple linear regression, pp 80 for multiple linear regression. Used to estimate sigma in standard error computations since sigma is usually unknown. Measures an absolute *lack* of fit.

*t*-statistic for the *j*th coefficient; pp 67. Measures the number of standard deviations that the coefficient is away from 0.

*p*-value;

pp 67. The probability of observing a value equal to |*t*| or larger, assuming the coefficient is actually zero. In other words the *p*-value is roughly the probability that the coefficient was at least as big as it was by chance when it was actually indistinguishable from zero. Used in hypothesis testing, small *p*-values imply that there is an association between the predictor and the target variable.

R2 pp 69. A Measure of fit of the model – the proportion of the target value (y) which can be explained by the model (higher is more explanation) Usually a number between 0 and 1, unless RSS > TSS (for really bad models!). Independent of scale of Y

Correlation(X,Y); pp 70. A measure of the linear relstionship between X and Y. Squared correlation give same value as R2 when in *simple* linear regression (1 feature). Can be used in feature exploration, but generally cannot be used to predict model fit in non-simple linear regression.

F-statistic; pp 75. Useful for a hypothesis test when determining if there is at least one non-zero coefficients. F >1 implies at least on non-zero coefficients

Leverage statistic for a point; pp 98. Captures how much influence the point has on the linear model fit. If the leverage statistic is above (*p*+1)/*n* then it is a high leverage point.

Variance inflation factor; pp 101. The variance of the coefficient in the full model divided by the variance of the coefficient if fit on its own. Used to detect multi-collinearity (collinearity between 3 or more variables, undetectable using only a correlation matrix). If VIF=1 then no collinearity. VIF>5 or 10 indicates a problem

From Chapter 3: Linear Regression Part 2: “Potential Problems”

For each of the problems listed below, explain a) what the problem is; b) how it would be detected in the model or the data and c) what should be done to resolve the problem.

1. Non-linearity of the response-predictor relationships
2. The underlying relationship between the data and the target variable is not linear. This can cause violation of some of the assumptions about applying a linear regressor as a model – which means the fit may be misleading. Any attempt to model such phenomenon with linear models will be suspect and give bad/misleading coefficients.
3. Detect using residual plots (plot estimated-y on the X-axis and the residual error in the estimated y on the Y-axis). A strong non-linear pattern in the residuals indicates that we have non-linearity.
4. To fix this problem, use non-linear transformations of the feature data such as log(X), sqrt(X) or X2, or use a non-linear model to fit the data.
5. Correlation of Error Terms
6. The error terms are correlated, meaning that information about *ei* tells us something about another error *ei*+1. This also means that our measured standard error is probably underestimating the true standard error – confidence intervals will be narrower than they should be. And this might cause us to say the feature is more significant than it really is – we are overconfident. Error terms are often correlated in time series data (for example, when the value of y at time = t+1 depends partially on the value of y at time = t)
7. When correlated errors exist in time-series data, residual plots of time series data predictions can reveal *tracking* where residuals appear to be adding or subtracting a small number (random walk) from the previous prediction. If the residuals are not correlated, then each value will appear independent of the previous value. This should be visible by taking cor(et,et+1) for all errors. If much above zero, errors may be correlated. Another method is to plot the error terms for t on the X axis and the residual for t+1 on the y axis. They should appear as a Gaussian blob if not correlated. If there is a linear pattern, they may be correlated.
8. Non-Constant Variance of error terms
9. Non-constant variance violates assumptions of the model which make the standard error, confidence intervals and hypothesis test results invalid.
10. Detection – make a plot of error (Y axis) as a function true *y* value (X axis) and look for differences in the width of the error boundaries. Example – a funnel shape is an indication of a violation of heteroscedasticity.
11. Solution: before fitting the model, transform the response *y* using log(*y*) or sqrt(*y*) to obtain a squished/compressed response variable that has less of a violation of the model assumptions. Then fit the model. When predicting, remember to use *ey* or *y*2 before reporting the final prediction. Another solutions is to use a weighted least squares method (see page 96)
12. Outliers
13. When a true value *yi* is far from its predicted value (might have been incorrect recording). Outliers can cause big changes to RSE which is the backbone of confidence intervals and hypothesis tests. Thus outliers cause misinterpretations of the fit.
14. Detect outliers with residual plots (y vs. (error=y-yhat)), possibly by studentizing (dividing each error first by its estimated standard error) to show how many standard deviations it is away from the expected position. Outliers might appear beyond 2 or 3 standard deviations in the studentized plot.
15. Solution – remove the outlier observation. Note that an outlier may indicate the model is incomplete (e.g. missing a feature)
16. High leverage points
17. Points which fall outside of the ‘bulk’ of feature values for the data (*X*1, *X*2,…*Xp*). These points have strong influence on linear models and can cause a misfit model. If the model changes a lot when the point is removed, it may be high leverage
18. Detection: compute the leverage statistic. If greatly exceeds (*p*+1)/*n* then might be high leverage.
19. Solution – similar to other outliers, consider removal, but not the caution about incomplete models
20. Collinearity
21. This is the problem when more than one feature are closely related, giving the impression that there is more info in the features than there really is. It also makes the coefficients fit between collinear features be sensitive to minor changes in the data. This makes it harder to distinguish individual contributions of the collinear features in the model (coefficients may be misleading). We may also fail to notice that collinear features are individually important (non-zero) in a model when doing a hypothesis test.
22. Detect pairwise collinearity with a correlation matrix. Any high values indicate a pair of highly collinear features. VIF can also be used to detect collinearity in more than 2 collinear features.
23. To fix collinearity, either drop one (or more) features or combine collinear features (mathematically) into a new feature (e.g. take the mean of the two features : new feature = (X­1+X2+…+Xk)/k