15.450 Final Exam

Professor Hui Chen

Spring 2020
Solutions

1. Explain the main difference(s) between AIC and BIC for model selection.

   Answer: The main difference between AIC and BIC is in their penalizing terms for model complexity. Suppose the model has T observations and p parameters, then the AIC penalizing term is $\frac{2}{T}p$, while BIC is $\frac{\ln T}{T}p$. Because usually $\ln T > 2$, BIC tends to select models with less parameters – that is, lower-order models in the time series setting.

2. Provide an example of a consistent but biased estimator, and an example of an inconsistent but unbiased estimator.

   Answer:
   Consistent but biased estimator: $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ is a consistent but biased estimator of $\sigma^2 = Var(x_i)$. Inconsistent but unbiased estimator: the first observation of the data is an unbiased estimator of the population mean, however it is inconsistent.

3. **True or false:** In an event study, one should try to specify the estimation window to be as long as possible in order to estimate the coefficients of the benchmark model accurately. (Please explain your answer.)

   Answer: False. Caveats of a longer estimation window include the model parameters are time varying, the data is non-stationary, and changes in the model itself (e.g., the factor structure changes).

4. **True or false:** One weakness of the value-at-risk (VaR) measure is that it requires returns to be normally distributed. (Please explain your answer.)

   Answer: False. Value-at-risk measure doesn't require returns to be normally distributed. As long as we have a model for the returns, we are able to estimate the model, forecast the distribution the returns based on the estimated model, and calculate the value-at-risk measure.

5. You are building a time-series model to predict the daily return volatility of the S&P500 ETF (ticker 'SPY').

   (a) How would you measure the daily return variance σ_t^2 of SPY using intra-day data (e.g. 5-minute returns)?

   Answer: Let's denote n equally spaced intra-day 5-minute log returns as $r_{t,i}, i = 1, \ldots, n$. The daily log return is $r_t^d = \sum_{i=1}^{n} r_{t,i}$. Assuming $r_{t,i}$ is i.i.d.,$\bar{r}_t \approx 0$ and n is large, then $\sigma_t^2 = Var(r_t^d) = nVar(r_{t,i}) \approx \sum_{i=1}^{n} r_{t,i}^2$. $\sum_{i=1}^{n} r_{t,i}^2$ is a consistent estimator of σ_t^2.

(b) Suppose you have constructed 10 potential features, x_{it} for $i = 1, \ldots, 10$, and you would like to construct a linear model to predict the (log) return variance on the next day:

$$\ln \sigma_{t+1}^2 = a + \sum_i b_i x_{it} + \epsilon_{t+1}. \tag{1}$$

Explain how you would build the model using best subset selection.

Answer: The best subset selection goes as the following.

(1) Start with the null model M_0 which contains no features.

(2) For $k = 1, 2, \ldots, 10$, fit all $\binom{10}{k}$ models that contain k features. Pick the best model among them, M_k, with the smallest RSS or largest R^2.

(3) Select a single best model from among $M_0, \ldots, M_{10}$ using cross-validation, AIC, BIC, or adjusted R^2...

(Optional) If we were to use cross-validation to select the best single model, we'd want to use the cross validation for time-series data as the following.

(1) Choose the minimum size of training set k.

(2) For $i = 1, 2, \ldots, T - k$, select the observation at $t = k + i$ as the test set. Use the observations from $t = 1$ to $t = k + i - 1$ as the training sample.

(3) Fit the model on the training sample and compute the prediction error for observation at $t = k + i$.

(4) Repeat Steps 2-3, and compute the overall average cross-validation error. The model with the lowest cross-validation error should be chosen.

6. Explain the meaning of the bias-variance trade-off in statistical learning. In particular, explain the meanings of the "bias" and "variance".

Answer: Suppose the true model of the data is $y = f(\mathbf{x})$ and the estimated model using the training data is $\hat{f}(\mathbf{x})$. The MSE of the model over new data $(y_0, \mathbf{x}_0)$ can be decomposed to three components,

$$MSE(\hat{f}(\mathbf{x}_0)) = \mathbb{E}[(y_0 - \hat{f}(\mathbf{x}_0))|\mathbf{x}_0] \tag{2}$$

$$= \sigma^2 + [f(\mathbf{x}_0) - \mathbb{E}(\hat{f}(\mathbf{x}_0))]^2 + \mathbb{E}[(\mathbb{E}(\hat{f}(\mathbf{x}_0)) - \hat{f}(\mathbf{x}_0))^2|\mathbf{x}_0] \tag{3}$$

$[f(\mathbf{x}_0) - \mathbb{E}(\hat{f}(\mathbf{x}_0))]^2$ is the Bias – i.e., error due to using the wrong model (f vs. $\hat{f}$). $\mathbb{E}[(\mathbb{E}(\hat{f}(\mathbf{x}_0)) - \hat{f}(\mathbf{x}_0))^2|\mathbf{x}_0]$ is the Variance – i.e., impact of the randomness of training sample on the fitted model $\hat{f}$. The Bias-variance trade-off refers to the phenomenon that simple models have high bias and low variance whereas more complex or sophisticated models have low bias and high variance. Having a bigger bias in exchange for smaller variance might be good for overall prediction accuracy.

7. When would we prefer to use LASSO over ridge regression to fit a linear model?

Answer: The L2-penalty in Ridge regression reduces the size of coefficients, but does not drive its elements to zero. Thus, it does not help with variable (model) selection. LASSO has advantage in doing variable selection by setting some of the coefficients to zero. When we have too many features to select from, we would prefer to use LASSO.

8. You are helping a fintech startup building a model to predict the default risk of mortgage borrowers. You have collected a large dataset on borrower characteristics (such as income and

leverage), market information (such as house prices and interest rates), and loan performances (default or non-default).

(a) Explain how you would apply the KNN method to predict whether a borrower will default in one year.

Answer: The KNN method goes as the following

(1) Identify K borrowers in the dataset that are closest to the borrower with attributes x_0, and denote this set of K borrowers as N_0.

(2) The conditional probability of the borrower default in one year is

$$\mathbb{P}(\text{default in one year}|x = x_0) = \frac{1}{K}\sum_{i \in N_0} \mathbb{I}(\text{borrower } i \text{ defaults in one year}) \quad (4)$$

(3) Pick a cutoff probability $\bar{p}$, and classify the borrower as "will default in one year" if $\mathbb{P}(\text{default in one year}|x = x_0) > \bar{p}$.

(4) Obtain the KNN decision boundary in the feature space.

(b) What advantages and disadvantages does KNN have compared to a logit model?

Answer:

Advantages: (1) Being a non-parametric method, KNN can capture the nonlinear relationship between features and the outcome, while logit model imposes a linear relationship which may not hold in reality.

Disadvantages: (1) Computation cost of KNN is higher. (2) Difficult to interpret the effect of each feature on the outcome in KNN.