Stat 627/427 (Statistical Machine Learning)

Homework # 6: Resampling Methods (Chap. 5 and notes)

Due March 6 by 12 noon on Blackboard. Or March 20, your choice. The Midterm is on March 7.

1. **(Jackknife and Bootstrap, continuing from the previous h/w)** One needs to estimate θ, the frequency of days with 0 traffic accidents on a certain highway. The data are collected. During 40 days, there are 26 days with 0 accidents, 10 days with 1 accident, and 4 days with 2 accidents.

   Statistician A estimates θ with a sample proportion $\hat{p} = 26/40 = 0.65$.

   Statistician B argues that this method does not distinguish between the days with 1 accident and the days with 2 accidents, losing some valuable information. She suggests to model the number of accidents X by a Poisson distribution with parameter λ. Then we have $\theta = \boldsymbol{P}\{X = 0\} = \exp(-\lambda)$. She estimates λ with $\hat{\lambda} = \bar{X}$. Then $\hat{\theta} = \exp(-\hat{\lambda})$. However, this estimator is biased, and last week, we reduced its bias by computing the jackknife estimator $\hat{\theta}_{JK}$.

   (a) Now we have three competing estimators - $\hat{p}$, $\hat{\theta}$, and $\hat{\theta}_{JK}$. Use bootstrap to estimate their standard deviations.

   (b) Compare our three estimators of θ according to their bias and standard error.

2. (**Bootstrap. Page 201, chap. 5, ≈#9**)
   We will now consider the `Boston` housing data set from the MASS library.

   (a) Based on this data set, provide an estimate for the population mean μ of `medv`, which is the median value of owner-occupied homes in \$1000s. Call this estimate $\hat{\mu}$.

   (b) Provide an estimate of the standard error of $\hat{\mu}$ (as we know, $\mathrm{Std}\,\bar{\mathrm{X}} = \sigma/\sqrt{\mathrm{n}}$).

   (c) Now estimate the standard error of $\hat{\mu}$ using the **bootstrap**. How does this compare to your answer from (b)?

   (d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for μ. A popular approximation is $\hat{\mu} \pm 2\,\widehat{\mathrm{Std}}(\hat{\mu})$. Compare it to the results obtained using an `R` command `t.test(Boston$medv)`.

   (e) Now, estimate M, the population *median* of `medv` with the sample median $\hat{M}$.

   (f) We now would like to estimate the standard error of $\hat{M}$, but unfortunately, there is no simple formula for computing the standard error of a sample median. Instead, estimate this standard error using the bootstrap.

The Midterm exam covers

- Regression (slopes, partial F-tests, lack of fit tests, categorical predictors, and interactions)
- Classification (logistic regression, LDA, QDA, and KNN)
- Cross-validation (validation set method, LOOCV, K-fold CV)
- Performance evaluation, prediction accuracy, and tuning (mean squared error and classification rate)