Stat 627/427 (Statistical Machine Learning)

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

Due February 27 by noon on Blackboard. Quiz #5 is on Feb 28.

1. **(Jackknife)** An acoustic studio needs to estimate the range of voice frequencies that an adult singer can produce. A sample of $n = 10$ recordings contains frequencies 102, 115, 127, 127, 162, 180, 184, 205, 239, 240.

   (a) Compute the jackknife estimator of the population highest frequency of a human voice.

   (b) Compute the jackknife estimator of the population lowest frequency of a human voice.

   (c) **(Stat-627 only)** Generalize the results. Assume a sample $X_1, \ldots, X_n$ of size n, where X_1, X_2 are the smallest two observations, and X_{n-1}, X_n are the largest two. Derive equations for the jackknife estimators of the population minimum and maximum.

   You can now compare your results in (a) and (b) with the natural range of human voice frequencies, and your bias correction in (c) with the formula that statisticians proposed in the article on the back during WWII.

2. **(Jackknife)** 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.

   (a) Use jackknife to estimate the bias of $\hat{\theta}$.

   (b) Compute the jackknife estimator $\hat{\theta}_{JK}$ based on $\hat{\theta}$.

   (c) Apply the jackknife method to estimate the bias of $\hat{p}$ and compute the jackknife estimator $\hat{p}_{JK}$. Explain the result.

   (d) Now, let us estimate the standard deviation of $\hat{p}$. The standard formula is

   $$s_{\hat{p}} = \widehat{\mathrm{Std}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

   Use jackknife to estimate the bias of $s_{\hat{p}}$. Propose an estimator with a smaller bias and calculate it from the given data. *(Knowledge of Poisson distribution is not needed to solve this problem.)*

   Can we compare the standard deviations of $\hat{p}$, $\hat{\theta}$, and $\hat{\theta}_{JK}$? For now, we have to table this, but next week, we'll learn bootstrap estimation of $\mathrm{Std}(\hat{\theta})$ *and* $\mathrm{Std}(\hat{\theta}_{\mathrm{JK}})$.

3. (**Cross-Validation. Pages 198-199, chap. 5, ≈#5**)
   Logistic regression is used to predict the probability of default using `income`, `balance`, and `student` on the `Default` data set in R package `ISLR`. Using

   (a) the validation set approach,
   (b) leave-one-out cross-validation, and
   (c) K-fold cross-validation for $K = 100$ and $K = 1000$,

   estimate the *test error* of this logistic regression model and decide whether it will be improved when the dummy variable `student` is excluded from this prediction.

Gavyn Davies does the maths

How a statistical formula won the war

Here is a story about mathematical deduction that I love, mainly because it is said to be true, and because it had an impact (albeit small) on the outcome of the second world war. It is the story of how a simple statistical formula successfully estimated the number of tanks the enemy was producing, at a time when this could not be directly observed by the allied spy network.

By 1941-42, the allies knew that US and even British tanks had been technically superior to German Panzer tanks in combat, but they were worried about the capabilities of the new marks IV and V. More troubling, they had really very little idea of how many tanks the enemy was capable of producing in a year. Without this information, they were unsure whether any invasion of the continent on the western front could succeed.

One solution was to ask intelligence to guess the number by secretly observing the output of German factories, or by trying to count tanks on the battlefield. Both the British and the Americans tried this, but they found that the estimates returned by intelligence were contradictory and unreliable. Therefore they asked statistical intelligence to see whether the accuracy of the estimates could be improved.

The statisticians had one key piece of information, which was the serial numbers on captured mark V tanks. The statisticians believed that the Germans, being Germans, had logically numbered their tanks in the order in which they were produced. And this deduction turned out to be right. It was enough to enable them to make an estimate of the total number of tanks that had been produced up to any given moment.

The basic idea was that the highest serial number among the captured tanks could be used to calculate the overall total. The German tanks were numbered as follows: **1**, **2**, **3** ... **N**, where **N** was the desired total number of tanks produced. Imagine that they had captured **five** tanks, with serial numbers **20**, **31**, **43**, **78** and**92**. They now had a sample of **five**, with a maximum serial number of **92**. Call the sample size **S** and the maximum serial number **M**. After some experimentation with other series, the statisticians reckoned that a good estimator of the number of tanks would probably be provided by the simple equation **$(M-1)(S+1)/S$**. In the example given, this translates to **$(92-1)(5+1)/5$**, which is equal to **109.2**. Therefore the estimate of tanks produced at that time would be **109**

By using this formula, statisticians reportedly estimated that the Germans produced **246** tanks per month between June 1940 and September 1942. At that time, standard intelligence estimates had believed the number was far, far higher, at around **1,400**. After the war, the allies captured German production records, showing that the true number of tanks produced in those **three** years was **245** per month, almost exactly what the statisticians had calculated, and less than **one fifth** of what standard intelligence had thought likely.

Emboldened, the allies attacked the western front in 1944 and overcame the Panzers on their way to Berlin. And so it was that statisticians won the war - in their own estimation, at any rate.