Bootstrap, Bagging, Random Forest and Adaboost Questions

Xintian Han

CDS, NYU

April 3, 2019

Question 1: What is random?

Which of the followings are random?

- statistic
- parameter
- point estimator
- sampling distribution
- standard error
- bias and variance of a point estimator
- bootstrap sample

Question 2: Bias and Variance

Suppose we have samples $X_1, \dots, X_N$ from $\text{Normal}(\mu, \sigma^2)$.

- What is the maximum likelihood estimator $\hat{\mu}, \hat{\sigma}^2$ of the parameter μ, σ^2?
- Are the maximum likelihood estimator unbiased? If not, how do we fix it?
- What is the sampling distribution of $\hat{\mu}$? What is the variance of the estimator $\hat{\mu}$?

Question 3: Bootstrap

Let $X_1, \ldots, X_{2n+1}$ be an i.i.d. sample from a distribution. To estimate the median of the distribution, you can compute the sample median of the data.

1. How do we compute an estimate of the variance of the sample median?
2. How do we compute an estimate of a 95% confidence interval for the median.

Question 4: Bagging and Random Forest

- A slide titled "Averaging Independent Prediction Functions" (there are multiple) states that $\mathrm{Var}(\hat{f}_{avg}(x)) = \frac{1}{B^2}\mathrm{Var}\left(\sum_{b=1}^{B}\hat{f}_b(x)\right) = \frac{1}{B}\mathrm{Var}\left(\hat{f}_1(x)\right)$. Justify each of the two equality signs.
- The above equality gives some intuition as to why bagging might reduce variance. But really, but situation is more complicated: the bootstrap samples used for bagging are not independent. Why not?
- If the variance of the individual predictors that we are bagging is σ^2, and the correlation between them is ρ^2, what is the variance of the bagged predictor?
- Bagging decision trees leads us to the highly popular random forests. However, to make bagging for decision trees work well, we need one more key idea. What is it?

Question 5: Adaboost-Concept Check[1]

Decide whether each of the statements below is true or false.

- If a weak classifier has a weighted error rate $\epsilon \leqslant 1/3$, it can only misclassify up to $1/3$ of the training points.
- The error rate of the ensemble classifier never increases from one round to the next.
- Adaboost accounts for outliers by lowering the weights of training points that are repeatedly misclassified.
- When you update weights, the training point with the smallest weight in the previous round will always increase in weight.
- Once a weak classifier is picked in a particular round, it will never be chosen in any subsequent round.

[1] From MIT exams

Question 6: Adaboost-Algorithm[2]

Consider building an ensemble of decision stumps Gm with the AdaBoost algorithm,

$$f(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

Figure displays a few labeled point in two dimensions as well as the first stump we have chosen. A stump predicts binary ± 1 values, and depends only on one coordinate value (the split point). The little arrow in the figure is the normal to the stump decision boundary indicating the positive side where the stump predicts $+1$. All the points start with uniform weights.

- Circle all the point(s) in Figure 1 whose weight will increase as a result of incorporating the first stump (the weight update due to the first stump).
- Draw in the same figure a possible stump that we could select at the next boosting iteration. You need to draw both the decision boundary and its positive orientation.
- Will the second stump receive higher coefficient in the ensemble than the first? In other words, will $\alpha_2 > \alpha_1$? Briefly explain your answer. (no calculation should be necessary).

[2] From CMU exams

Question 6:Adaboost-Algorithm