Permutation Tests and Bootstrapping

Uri Shaham

March 4, 2024

1 Bootstrapping

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{P}$ be iid variables and for each i, let x_i be a realization of X_i. Let $X = (X_1, \ldots, X_n)$ $x = (x_1, \ldots, x_n)$ Let $S(X)$ be a statistic, with observed value $s(x)$. We would like to estimate the distribution of $S(X)$, based on the observed data.

Definition 1.1 (Empirical distribution)**.** *Given the above setting, the empirical distribution $\hat{\mathcal{P}}$ is obtained by putting $\frac{1}{n}$ mass on each x_i.*

Definition 1.2 (Bootstrap sample)**.** *A bootstrap sample of size m is $X^* = (X_1^*, \ldots, X_n^*)$, where $X_1^* \overset{\text{iid}}{\sim} \hat{\mathcal{P}}$.*

Given a bootstrap sample X^* we can approximate the distribution of $S(X)$ via that of $S * (X^*)$, which is the distribution of the statistic induced from X^*.

Example 1.3. *Let $\mathcal{P} = Ber(\theta)$. We know that $\bar{X}$ is an unbiased estimate of θ. However, we would like to understand the distribution of the error $S(X) := \bar{X} - \theta$. Having observed x, $\hat{\mathcal{P}} = Ber(\bar{x}$. Let X^* be a bootstrap sample from $\hat{\mathcal{P}}$, and let $S^*(X^*) = X^* - \bar{x}$. Then $n\bar{X}^* \sim Bin(n, \bar{x})$, so $\mathbb{E}[S^*(X^*)] = 0$, and $\text{Var}(S^*(X^*)) = \frac{\bar{x}(1-\bar{x})}{n}$.*

Example 1.4. *Let $\mathcal{P}$ be a uniform distribution over the heights of the worldwide population, and we are interested in estimating is mean μ. Since we cannot measure them all, we obtain a sample $x = (x_1, \ldots, x_n)$ of heights, which can be thought of as realizations of $X_1, \ldots, X_n$, n iid copies of X. Then $\bar{x}\frac{1}{n}\sum_i x_i$ is an estimate of μ, however this is only a single point estimate. We would like to know how much $S(X) = \bar{X}$ varies, or put another way, to infer about the distribution of $S(X) = \bar{X}$. For this we can use Monte Carlo sampling: create k bootstrap samples $X_1^*, \ldots X_k^*$. Each such sample (of size(n) would give is an estimate $S(\bar{x}_k^*)$ of , and we can create a histogram of these values and have it as an approximation of the distribution of $S(X) = \bar{X}$.*

2 Permutation Tests

Statistical hypothesis test is a procedure where data is examined in order to conclude whether to accept a given hypothesis about some parameter of interest. A permutation test is a non-parametric procedure (i.e., it doesn't assume any parametric form of the population), and is a highly useful procedure for hypothesis testing. We will demonstrate the idea of hypothesis testing through example, taken from `https://www.jwilber.me/permutationtest/`. Let $X_A \sim \mathcal{P}_A, X_B \sim \mathcal{P}_B$ be two random variables with means μ_A, μ_B. We collect n_A and n_B iid samples from each and we want to conclude whether$\mu_A = \mu_B$. Specifically, suppose the we want to examine the efficiency of a new shampoo to the quality of alpaca

wool. We have a population of alpacas, and we randomly divide them to two groups, A and B, and treat only the B alpacas with the shampoo.

We can formalize the hypotheses as:

- $H_0 : \mu_A = \mu_B$
- $H_1 : \mu_A < \mu_B$

We know that sample averages are unbiased estimates for population means, hence it is natural to define the test statistic as $S(X) = \bar{X}_B - \bar{X}_A$, with observed value $S(x) = \bar{x}_B - \bar{x}_A$, obtained after treating the B alpacas with the new shampoo. This will give as a single value, say, 1.7. However, we don't know how to interpret this value, as in order to do so, we need to know the distribution of the test statistic under the null hypothesis. In a parametric test, we can, for example, assume that $\mathcal{P}_A = \mathcal{N}(\mu_A, \sigma)$, and $\mathcal{P}_B = \mathcal{N}(\mu_B, \sigma)$, in which case a student t test can be performed. However, often times such assumption is problematic, and we would like to avoid it. In such cases, a non-parametric is preferable. The idea of permutation test is simple: we can use bootstrapping to approximate the distribution of $S(X)$ using empirical distribution. Specifically, we can permute our alpaca population and divide them randomly to two new groups A' and B;, and compute the test statistic $s(x')$. Clearly, since the partition is random, $\mu_{A'} = \mu_{B'}$, so the null hypothesis holds in this simulation. Repeating this a large number of times (say, 10,000) and collecting the values of the test statistic from all permutation's, will give us a histogram approximating the distribution of the test statistic under the null hypothesis.

Given the true statistic $s(x)$, we can use the histogram to $1 - F(S = s)$, where F is the cumulative distribution function of S under the null hypothesis. This is simply an integral of the histogram. This is our p-value, which is the probability, under the null hypothesis, to get a statistic value at least as extreme as what we got. The smaller the p-value is, the more likely it is that H_0 is false. If we define a rejection area of the form (a, ∞), whose total probability is α (say, $\alpha = 5\%$), the meaning of the p-value is the minimal α needed to reject H_0 (i.e., when $a = s$).

3 Aside: Adjusting classifier outputs to change in class prior probabilities

Let (X, Y) be pair of random variables, where X can be thought of as data and Y as (symbolic) label. Then the joint probability is $P(X, Y) = P(Y)P(X|Y)$. We call $p(Y)$ the prior probability. The prior can be easily estimated from training data by $\hat{P}(Y = i) = \frac{n_i}{n}$, where n_i is the number of data points from class i and n is the total size of the training data. Suppose we have iid training samples from $P(X, Y)$, which we used to train a wonderful classifier. Now the classifier is ready to be deployed. Unfortunately, the data distribution in the real world is $P'(X, Y) = P'(Y)P(X|Y)$, i.e., we assume that the conditional elements $P(X|Y)$ are unchanged, however the prior component is changed (e.g., the proportion of dog to cats may be different in the real world than in our training data). This is a case of domain adaptation that has an elegant and simple solution.

Our trained classifier outputs $P(Y|X)$. We also know (or can easily estimate) $P(Y)$. What we want is the true posterior $P'(Y|X)$.

3.1 Known $P'(Y)$

For a start, suppose we know $P'(Y)$. Our goal is now to express $P'(Y|X)$ as a function of $P(Y|X), P(Y), P'(Y)$. Using Bayes theorem, we have

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}.$$

Since we assume that $P(X|Y) = P'(X|Y)$ (i.e., $P(X|Y)$ remains unchanged, we have

$$\frac{P(X)P(Y|X)}{P(Y)} = \frac{P'(X)P'(Y|X)}{P'(Y)},$$

i.e.,

$$P'(Y|X) = \frac{P(X)P'(Y)}{P'(X)P(Y)}P(Y|X).$$

Since we know $P(Y|X), P(Y)$ and $'P(Y)$, what remains is to estimate $\frac{P(X)}{P'(X)}$. To do that, note that $\sum_i P'(Y=i|X) = 1$, therefore $\frac{P(X)}{P'(X)}\sum_i \frac{P'(Y=i)}{P(Y=i)}P(Y|X) = 1$, and

$$\frac{P(X)}{P'(X)} = \left(\sum_i \frac{P'(Y=i)}{P(Y=i)}P(Y|X)\right)^{-1}.$$

Altogether, this gives us

$$P'(Y|X) = \frac{\frac{P'(Y)}{P(Y)}P(Y|X)}{\sum_i \frac{P'(Y=i)}{P(Y=i)}P(Y|X)}.$$

3.2 Unknown $P'(Y)$

The more realistic case is when we actually don't know $P'(Y)$. One way to solve this case is using what's known as the "confusion matrix method". For k classes, confusion matrix is a $k \times k$ matrix, whose i, j entry is the probability to assign label j to an example from class i, that is

$$\Pr(\text{decision is j}|Y = i).$$

We can estimate the confusion matrix from the training data. Note that our assumpsion that $P'(X|Y) = P(X|Y)$ implies that

$$P'(\text{decision is j}) = \sum_i P(\text{decision is j}|Y = i)P'(Y = i).$$

The $P'(\text{decision is j})$ terms can be estimated from the test data. Therefore we obtain a system of k equations in k unknowns (the $P'(Y = i)$ terms). Solving this system gives us the $P'(Y = i)$'s, and now we can apply the formula from the previous subsection.

Homework

1. Suggest a test statistic and a permutation test-based method to evaluate the quality of generative models (e.g., a model that generates deep fake face images) . Discuss its advantages and disadvantages.

2. Say we have a test statistic S with observed value and we define our p-value as $P = F(S)$, where F is the cumulative distribution function of S under the null hypothesis. Show that P is uniformly distributed under the null hypothesis.