1 Introduction

- Most of Machine Learning (ML) algorithms were proposed and hand-tuned to solve very specific problems, e.g., face recognition, text analysis, playing chess, etc. In this sense, ML is sort of "statistical engineering."
- ML borrows a lot from statistics, but statistical properties of many commonly used ML algorithms are not so well understood. Broadly speaking, "good performance" of the algorithm is not always theoretically guaranteed and it may be hard to make formal probabilistic statements about it.
- Distinctive features of ML:
  - Clear metric of success: out-of-sample prediction accuracy;
  - Algorithmic approach: minimum guidance required from the user;
  - Sparsity/regularization/dimensionality reduction;
  - Model selection/adaptivity;
- Examples in economics:
  - Demand analysis: many products/attributes (Bajari/Nekipelov/Ryan/Yang)
  - Dynamic games: large state space (Aguirregabiria/Conrad-Wexler/Ryan)
  - Network models: missing links data (Manresa)
  - New types of data (e.g. text data) (Gentzkow/Shapiro/Taddy)
  - Lasso (theory: Bickel/Ritov/Tsybakov, Belloni/Chen/Chernozhukov/Hansen)
  - Matching data sets (Feigenbaum)
  - Causal inference: estimating heterogeneous treatment effects (Athey/Wager)
  - Structural models via indirect inference (Kaji/Manresa/Poulliot)
  - Double ML (Chernozhukov/Chetverikov/Demirer/Duflo/Hansen/Newey/Robins)

2 A Review of Statistics

The purpose of the first lecture is to make sure that everyone is on the same page. You should be familiar with the concepts already, but the exposition below may be more formal than what you've seen before.

2.1 Estimation. Asymptotic Approximation.

We will start with some definitions. A *random sample*, $X_1, \ldots, X_n \sim P$, is a collection of independently and identically distributed random variables (or vectors) with marginal distribution P. It will be useful to denote $X_1^n = (X_1, \ldots, X_n)$.

An estimand $\theta = \theta(P)$ is an unknown parameter of interest, which typically represents some feature of the distribution P. The goal of statistical analysis is to estimate this parameter and conduct inference about it, that is, to test hypotheses and construct confidence intervals. In this course, whenever we talk about the "unknown parameter" or "parameter of interest," think of it as of a particular feature of the distribution of the data.

An estimator $\hat{\theta}_n$ is any function of the data. Since the data is sampled at random, all estimators are random variables. Therefore, we will often make statements about their distributions, means, and variances.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution P. One parameter of interest is the mean $\theta(P) = \mathbb{E}_P[X]$. To estimate the mean, we typically use the sample average:

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This $\hat{\theta}_n$ is a random variable, and, for example,

$$\mathbb{E}_P[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P[X_i] = \mathbb{E}_P[X_i]$$

and

$$\mathbb{V}ar_P(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}ar_P(X_i) = \frac{\mathbb{V}ar_P(X_i)}{n}.$$

Below, we will say a lot more about the properties of this estimator.

Another parameter of interest could be $\theta(P) = \mathbb{V}ar_P(X)$. A reasonable estimator is

$$\hat{\theta} = \hat{\sigma}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

This estimator is motivated by the variance formula

$$\mathbb{V}ar(X) = \mathbb{E}[(X - \mathbb{E}[X])^2],$$

where we have replaced all expectations with sample averages. ■

Replacing expectations with sample averages is a common trick called the *analogy principle* or the *plug-in principle*. It is based on the idea that the underlying distribution of the data P can be estimated by the distribution $\hat{P}_n$ placing a probability mass $1/n$ on each of the observations X_i. Then, to estimate $\theta(P) = \mathbb{E}_P[g(X)]$, we "plug in" $\hat{P}_n$ obtaining $\theta(\hat{P}_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)$ (this is exactly how we calculate the mean for the discrete distribution $\hat{P}_n$). If this seems a little too abstract, just remember to replace expectations with sample averages.

Next, we need the following concept. A sequence W_n of random variables *converges in probability* to a constant c if for any $\varepsilon > 0$,

$$P(d(W_n, c) > \varepsilon) \to 0 \quad \text{as } n \to \infty$$

where $d(W_n, c)$ represents the distance between W_n and c.[1] In words, it means that the sequence W_n gets arbitrarily close to, or "trapped around" c with high probability for large n. Indeed, for some fixed $\varepsilon > 0$, the event $\{d(W_n, c) > \varepsilon\}$ means that W_n "far" from c, and we require that the probability of this event approaches zero as n increases. Convergence in probability will be denoted by $W_n \to_p c$.

For any parameter of interest, one can come up with many different estimators. What does it mean to have a "reasonable" estimator? The key property that we will require from all estimators is *consistency*. An estimator $\hat{\theta}_n$ is *consistent* for a parameter θ if, for any $\varepsilon > 0$,

$$P\left(d(\hat{\theta}_n, \theta) > \varepsilon\right) \to 0 \quad \text{as } n \to \infty,$$

or $\hat{\theta}_n \to_p \theta$. Consistency is a global property: an estimator that in principle may end up anywhere in the parameter space converges to the true value. Put another way, if an estimator is inconsistent, it is not estimating what we want.

[1] If $x, y \in \mathbb{R}^d$, we can think of $d(x, y) = \left(\sum_{j=1}^d (x_j - y_j)^2\right)^{1/2}$, but any other distance function will do.

The following results help establish consistency in practice.

Theorem 1 (Law of Large Numbers)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution P with $\mathbb{E}_P[|X|] < \infty$. Denote $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then,*

$$\bar{X}_n \quad \to_p \quad \mathbb{E}_P[X].$$

In words, the Law of Large Numbers (LLN) states that sample averages get close to expectations with high probability, as $n \to \infty$. This is exactly what motivates the analogy principle. We will see its applications repeatedly throughout the course.

Theorem 2 (Continuous Mapping Theorem)**.** *If $W_n \to_p c$ and a function $g(w)$ is continuous at c, then $g(W_n) \to_p g(c)$.*

This powerful property (CMT for short), combined with the LLN, allows to show consistency of very complicated estimators as long as they can be expressed as continuous functions of moments. The requirement of continuity is essential: if $W_n \to_p c$ but the function $g(w)$ has a jump at c, then even though W_n is close to c with high probability, $g(W_n)$ will still be "far" from $g(c)$ unless $W_n = c$ holds exactly.

Example 2.

1. Suppose a random sample $X_1, \ldots, X_n$ is given. To estimate $\mathbb{E}_P[X^2]$ consistently, one can use $\frac{1}{n}\sum_{i=1}^n X_i^2$. To see this, apply the LLN to $Y_i = X_i^2$.

2. To estimate $(\mathbb{E}_P[X])^2$ consistently, one can use $(\bar{X}_n)^2$. To see this, apply the CMT to $W_n = \bar{X}_n$ and $g(w) = w^2$.

3. Let $X_i = (Y_i, Z_i)$ and consider a parameter

$$\beta(P) = \frac{\mathbb{E}_P[Y_i Z_i]}{\mathbb{E}_P[Z_i^2]}.$$

   The numerator can be consistently estimated by $\frac{1}{n}\sum_i Y_i Z_i$, and the denominator — by $\frac{1}{n}\sum_i Z_i^2$. Since the function $g(a, b) = \frac{a}{b}$ is continuous for $b \neq 0$, it follows from the CMT that

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n Y_i Z_i}{\frac{1}{n} Z_i^2}$$

   is a consistent estimator for $\beta(P)$.

■

The above discussion concerns *point estimation*, which doesn't convey any information about uncertainty. For example, suppose that we ask 100 random people on the streets of Chicago about their annual income and obtain the sample average of 58,000$ per year. It is not clear how close this number is to the true mean: the latter could be equal to 57,200$, which is pretty close, or 70,000$, which seems far off. The difference between the reported sample average and the true mean is due to the sampling uncertainty: if we had performed the same experiment on a different day, we would have obtained a different sample average, which could have been be closer to or further away from the true value. Therefore, it would be good to have some tools to *quantify the uncertainty*. This is why we are interested in inference, i.e., testing hypothesis and constructing confidence intervals.

To proceed, we will need the following definition. A sequence of random variables Z_n *converges in distribution* to a random variable Z with a continuous distribution if

$$P(Z_n \in [a,b]) \to P(Z \in [a,b]) \quad \text{as } n \to \infty$$

for all $a \leqslant b$. Convergence in distribution is denoted by $Z_n \to_d Z$.

This definition is the basis for *asymptotic approximation*. Let Z_n be some function of the data. Since the distribution of the data is unknown, we cannot say much about the distribution of Z_n for a fixed n. However, using the above definition, we can approximate the distribution of Z_n by the distribution of Z for large n. Typically, Z will be a Normally distributed random variable so that we can compute or estimate the probabilities of the form $P(Z \in [a,b])$. Then, we will write approximately $P(Z_n \in [a,b]) \approx P(Z \in [a,b])$ and proceed to construct asymptotically valid tests or confidence intervals.

To develop asymptotic approximations, we will use the following results.

Theorem 3 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution P with $\mathbb{E}_P[X] = \mu$ and $\mathbb{V}ar_P[X] = \sigma^2 < \infty$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then:*

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \quad \to_d \quad Z \sim N(0,1).$$

In words, the Central Limit Theorem (CLT) states that sample averages of i.i.d. data are approximately Normally distributed: $\bar{X}_n \overset{appr}{\sim} N(\mu, \sigma^2/\sqrt{n})$. Informally, note that as n increases, the variance in this approximation converges to zero, implying that $\bar{X}_n$ converges to μ in some sense, which is in line with the LLN. This result can be generalized in a number of ways (for vectors, for independent but not identically distributed data, for certain time series), but the key is to have independent or "almost independent" observations.

We will also rely on the following results.

Theorem 4 (Continuous Mapping Theorem)**.** *If $Z_n \to_d Z$ and g is a continuous function, then $g(Z_n) \to_d g(Z)$.*

This property, combined with the CLT, allows to show convergence in distribution of statistics, which combine multiple sample averages in a potentially non-linear way.

Theorem 5 (Slutsky's Theorem)**.** *If $W_n \to_p c$, $Z_n \to_d Z$, and $g(w, z)$ is a continuous function, then $g(W_n, Z_n) \to_d g(c, Z)$.*

This theorem allows to replace unknown values (e.g. the standard deviation in the CLT formula) with their consistent estimators when establishing convergence in distribution.

Theorem 6 (Delta Method)**.** *If $\sqrt{n}(\hat{\theta}_n - \theta) \to_d Z$ and f is differentiable at θ, then*

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta)) \quad \to_d \quad f'(\theta) \cdot Z.$$

In particular, if $Z \sim N(0, \sigma^2)$, then $\sqrt{n}(f(\hat{\theta}_n) - f(\theta)) \to_d N(0, f'(\theta)^2\sigma^2)$.

This theorem gives asymptotic approximations for differentiable functions of estimators. To illustrate, consider the following examples.

Example 3. Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution P with $\mathbb{E}_P[X] = \mu$ and $\mathbb{V}ar_P[X] = \sigma^2$. By the CLT,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \quad \to_d \quad Z \sim N(0, 1).$$

In practice, to construct a confidence interval or a test for μ, we cannot use this result directly because σ is unknown. In the home assignment, you will show that the estimator

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is consistent for σ^2, that is, $\hat{\sigma}_n^2 \to_p \sigma^2$. Then, the CMT implies that $\hat{\sigma}_n \to_p \sigma$, since $g(x) = \sqrt{x}$ is a continuous function. Therefore, by Slutsky's theorem,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n} = \underbrace{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}_{\to_d Z} \cdot \underbrace{\frac{\sigma}{\hat{\sigma}_n}}_{\to_p 1} \quad \to_d \quad Z \sim N(0, 1)$$

since $g(w, z) = zw$ is a continuous function.

By the CMT (Theorem 4), we can also conclude that:

$$\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n}\right)^2 \quad \to_d \quad Z^2 \sim \chi^2(1).$$

On the other hand, by the Delta-Method applied with $f(x) = x^2$, we have:

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \quad \to_d \quad (2\mu) \cdot X,$$

where $X \sim N(0, \sigma^2)$. Therefore, $\sqrt{n}(\bar{X}_n^2 - \mu^2) \to_d N(0, 4\mu^2\sigma^2)$.

■

2.2 Inference

2.2.1 Hypothesis testing

Any statement about an unknown parameter is a hypothesis. For example, denoting $\theta = \mathbb{E}_P[X]$, we can hypothesize $\theta = 2$, or $\theta \geqslant 0$. The data $X_1, \ldots, X_n$ will provide evidence in favor of the hypothesis or against it. Hypotheses are formulated in pairs: the null (e.g. $\theta = 2$) and the alternative (e.g. $\theta \neq 2$).

Having collected the data, we can test the null hypothesis, that is, check if the data supports or contradicts it. Since the data is random, we will sometimes make mistakes. Rejecting the null hypothesis when it is, in reality, true, is called *type-I error*. Rejecting the alternative hypothesis when it's true is called *type-II error*. The tests are typically constructed to control the probability of type-I error (e.g. $\alpha = 0.05$). This probability is also called the *level* of the test.

Suppose we're interested in testing $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. The goal is to construct a test for which the probability of type-I error is equal to some pre-specified level α. Since we're relying on the asymptotic approximation, the formal statement looks like:

$$\lim_{n\to\infty} P(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$$

We start by constructing an estimator $\hat{\theta}_n$ such that:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \quad \to_d \quad N(0, \sigma^2). \tag{1}$$

Then, we obtain a consistent estimator for the asymptotic variance, $\hat{\sigma}_n^2$, and work with a test

statistic

$$Z_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\sigma}_n} \quad \to_d \quad N(0,1),$$

arguing as in Example 3. Letting $z_{\alpha/2}$ denote the $\alpha/2$ upper quantile of $N(0,1)$, we use the rule:

$$\text{Reject } H_0 \text{ if } \quad |Z_n| > z_{\alpha/2}.$$

Why is this a suitable test? Recall that the goal is to control the probability of type-I error: $P(\text{reject } H_0 \mid H_0 \text{ is true})$. We have:

$$\begin{aligned}
\lim_{n\to\infty} P(\text{reject } H_0 \mid H_0 \text{ is true}) &= \lim_{n\to\infty} P(|Z_n| > z_{\alpha/2} \mid \theta_0 \text{ is the true value}) \\
&= P(|Z| > z_{\alpha/2}) \\
&= \alpha,
\end{aligned}$$

where the second line is justified by out asymptotic approximation. Therefore, this test controls the probability of type-I error asymptotically.

2.2.2 Confidence Intervals

A confidence interval is a data-dependent interval, denoted $CI_{1-\alpha,n}$, that covers the true parameter value θ with some pre-specified probability $1-\alpha$. Since we're relying on an asymptotic approximation, the formal statement will look like:

$$\lim_{n\to\infty} P(\theta \in CI_{1-\alpha,n}) = 1 - \alpha.$$

To construct a confidence interval for a parameter θ, we start by constructing an estimator:[2]

$$\sqrt{n}(\hat{\theta}_n - \theta) \quad \to_d \quad N(0, \sigma^2).$$

Arguing as before,

$$Z_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\hat{\sigma}_n} \quad \to_d \quad Z \sim N(0,1)$$

Therefore,

$$\lim_{n\to\infty} P(-z_{-\alpha/2} \leqslant Z_n \leqslant z_{\alpha/2}) = P(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}) = 1 - \alpha.$$

[2] In Equation (1) we had θ_0 because under H_0, $\theta = \theta_0$. Here, there is no H_0, so the true parameter value is denoted by θ.

Plugging-in Z_n and re-arranging,

$$\lim_{n\to\infty} P\left(\hat{\theta}_n - z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}} \leqslant \theta \leqslant \hat{\theta}_n + z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right) = 1-\alpha.$$

Therefore,

$$CI_{1-\alpha,n} = \left[\hat{\theta}_n - z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}},\ \ \hat{\theta}_n + z_{\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$$

is the sought confidence interval.

2.3 Homework

1. Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution P with $\mu = \mathbb{E}_P[X]$ and $\sigma^2 = \mathbb{V}ar_P(X)$. Show that

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

   is a consistent estimator for σ^2. (Hint: expand the square and simplify before applying any theorems).

2. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution $U[0,\theta]$. Show that $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$ is a consistent estimator for θ, that is, for any $\varepsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \varepsilon) \to 0$$

   as $n \to \infty$. Extra question: what is the asymptotic distribution of $n(\hat{\theta}_n - \theta)$? (Hint: derive the CDF and take limit as $n \to \infty$).

3. Airlines often overbook flights to maximize profit. Last year, I flew about 50 times and about 5 of those flights were overbooked. Let $X_i = 1$ if the flight it overbooked, and $X_i = 0$ otherwise. Given $X_1, \ldots, X_{50}$, how would you estimate the probability of a flight being overbooked, $p = P(X = 1)$? Denoting your estimator by $\hat{p}_n$, what is the asymptotic distribution of $\sqrt{n}(\hat{p}_n - p)$? Based on this asymptotic approximation, how would you construct a 95% confidence interval for this parameter? Extra question: letting σ^2 denote the asymptotic variance of $\hat{p}_n$ and $\hat{\sigma}_n^2$ denote an consistent estimator for it, what is the asymptotic distribution of $\sqrt{n}(\hat{\sigma}_n - \sigma)$?

4. Write a Monte Carlo simulation to check coverage of the confidence interval for the mean with known variance and with estimated variance. For each simulation:

(1) Draw $X_1, \ldots, X_n$ from a distribution of your choice, for which you know the mean and the variance.

(2) Construct two confidence intervals: with known standard deviation and with the estimated one.

(3) Check if each of them covers the true value of the mean and save the result.

Perform 1000 simulations and report how often each of the intervals covers the true value. Repeat this exercise for $n = 30, 100$, and 500. Discuss the results.