Introduction to Data Science - 1MS041

Benny Avelin

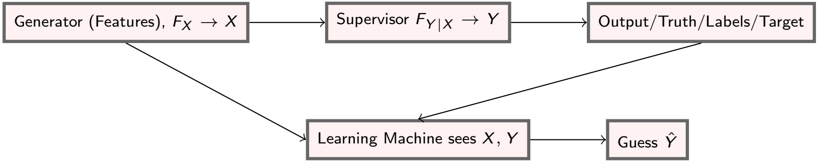
Department of Mathematics

HT 2023

Recall from last time

Setup

1. The generator of the data G
2. The supervisor S
3. The learning machine LM.



Recall from last time

- Statistical model (Our assumptions of the truth)

Recall from last time

- Statistical model (Our assumptions of the truth)
- The model space $\mathcal{M}$, what the learning machine searches in.

Recall from last time

- Statistical model (Our assumptions of the truth)
- The model space $\mathcal{M}$, what the learning machine searches in.
- The loss function L measuring the performance of a function $g \in \mathcal{M}$ w.r.t data.

Recall from last time

- Statistical model (Our assumptions of the truth)
- The model space $\mathcal{M}$, what the learning machine searches in.
- The loss function L measuring the performance of a function $g \in \mathcal{M}$ w.r.t data.
- The risk which is expected loss.

Recall from last time

- Statistical model (Our assumptions of the truth)
- The model space $\mathcal{M}$, what the learning machine searches in.
- The loss function L measuring the performance of a function $g \in \mathcal{M}$ w.r.t data.
- The risk which is expected loss.
- The main objective of the learning machine is to find $\hat{g} \in \mathcal{M}$ that minimizes risk.

Recall from last time

- We talked about the following learning problems
  - Find f
  - Regression
  - Pattern recognition

Recall from last time

- We talked about the following learning problems
  - Find f
  - Regression
  - Pattern recognition
- We defined the regression function

$$r(X) = \mathbb{E}[Y \mid X]$$

which is the target to hit with Regression.

Estimation

Recap

Up to now we have brushed upon a general construction, namely that of an estimator.

Estimation

Recap

Up to now we have brushed upon a general construction, namely that of an estimator.

Definition (Data)

Assume that $X = (X_1, \ldots, X_n)$ is a sequence of $\mathbb{R}^m$ valued random variables taking values in the **data space** $\mathbb{X}$:

Estimation

Recap

Up to now we have brushed upon a general construction, namely that of an estimator.

Definition (Data)

Assume that $X = (X_1, \ldots, X_n)$ is a sequence of $\mathbb{R}^m$ valued random variables taking values in the **data space** $\mathbb{X}$:

$$X(\omega) : \Omega \to \mathbb{X} \ .$$

Sometimes we call X a **Dataset** or just **Data**.

Estimation

Recap

Up to now we have brushed upon a general construction, namely that of an estimator.

Definition (Data)

Assume that $X = (X_1, \ldots, X_n)$ is a sequence of $\mathbb{R}^m$ valued random variables taking values in the **data space** $\mathbb{X}$:

$$X(\omega) : \Omega \to \mathbb{X} \; .$$

Sometimes we call X a **Dataset** or just **Data**. The realisation of the RV X when an experiment is performed is the observation or **data/dataset** $x \in \mathbb{X}$.

Estimation

Definition (Statistic)

Consider a data space $\mathbb{X}$. A **statistic** T is a function on the data space:

$$T : \mathbb{X} \to \mathbb{T} \ .$$

Estimation

Definition (Statistic)

Consider a data space $\mathbb{X}$. A **statistic** T is a function on the data space:

$$T : \mathbb{X} \to \mathbb{T} \ .$$

Examples of estimators and spaces

- $\mathbb{X} = \mathbb{R}^n$, and $T(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$. In this case $\mathbb{T} = \mathbb{R}$.

Estimation

Definition (Statistic)

Consider a data space $\mathbb{X}$. A **statistic** T is a function on the data space:

$$T : \mathbb{X} \to \mathbb{T} \ .$$

Examples of estimators and spaces

- $\mathbb{X} = \mathbb{R}^n$, and $T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$. In this case $\mathbb{T} = \mathbb{R}$.
- $\mathbb{X} = (\mathbb{R}^2)^{\otimes n}$, $X = ((X_1, Y_1), \ldots, (X_n, Y_n))$. Now consider the linear regression problem, let $T[x] = g^*[x]$ be the best fitting linear function on the dataset $x \in \mathbb{X}$. In this case $\mathbb{T}$ is the set of all linear functions.

See simulations:

More examples

- The empirical risk, of a function $g \in \mathcal{M}$

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$

More examples

- The empirical risk, of a function $g \in \mathcal{M}$

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$

- Let $X_{train} = ((X_1, Y_1), \dots, (X_n, Y_n))$ be training Data, and consider

$$g^* := \arg\min_{g \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$

  Now, consider a new Dataset
  $X_{test} = ((X_{n+1}, Y_{n+1}), \dots, (X_{n+m}, Y_{n+m}))$, then look at

$$\frac{1}{m} \sum_{i=n+1}^{m} L(g^*(X_i), Y_i)$$

  Given g^* the above is an estimator w.r.t. the X_{test} Dataset.

Terminology

Definition

Consider the statistical model

$$\mathcal{E} = \{F(x;\lambda) : \mathbb{X} \to [0,1] : \lambda \in \Lambda,\ F \text{ is a DF}\}$$

Let a parameter map be given $\theta : \Lambda \to \mathbf{\Theta}$. Consider the Data $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} F(\cdot;\lambda^*) \in \mathcal{E}$ be $\mathbb{R}^m$-valued RVs.

Terminology

Definition

Consider the statistical model

$$\mathcal{E} = \{F(x;\lambda) : \mathbb{X} \to [0,1] : \lambda \in \Lambda,\ F \text{ is a DF}\}$$

Let a parameter map be given $\theta : \Lambda \to \mathbf{\Theta}$. Consider the Data $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} F(\cdot;\lambda^*) \in \mathcal{E}$ be $\mathbb{R}^m$-valued RVs. A **point estimator** of $\theta^* := \theta(\lambda^*) \in \mathbf{\Theta}$ is a statistic, i.e.

$$\hat{\Theta} : \mathbb{X} \to \mathbf{\Theta},$$

sometimes we denote it as $\hat{\Theta}_n$ to highlight that it depends on n values.

Terminology

Definition

Consider the statistical model

$$\mathcal{E} = \{F(x;\lambda) : \mathbb{X} \to [0,1] : \lambda \in \Lambda,\ F \text{ is a DF}\}$$

Let a parameter map be given $\theta : \Lambda \to \mathbf{\Theta}$. Consider the Data $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} F(\cdot;\lambda^*) \in \mathcal{E}$ be $\mathbb{R}^m$-valued RVs. A **point estimator** of $\theta^* := \theta(\lambda^*) \in \mathbf{\Theta}$ is a statistic, i.e.

$$\hat{\Theta} : \mathbb{X} \to \mathbf{\Theta},$$

sometimes we denote it as $\hat{\Theta}_n$ to highlight that it depends on n values. The bias of an estimator $\widehat{\Theta}_n$ of $\theta^* \in \mathbf{\Theta}$ is:

$$\text{bias}(\widehat{\Theta}_n(X)) := \mathbb{E}(\widehat{\Theta}_n(X)) - \theta^* = \int \widehat{\Theta}_n(x)\, dF(x;\lambda^*) - \theta(\lambda^*)\ . \quad (1)$$

Example: parametric model

$$\mathcal{N} = \{N(0, \sigma^2) : \sigma \in (0, \infty)\}$$

Example: parametric model

$$\mathcal{N} = \{N(0, \sigma^2) : \sigma \in (0, \infty)\}$$

- Here $\Lambda = (0, \infty)$.

Example: parametric model

$$\mathcal{N} = \{N(0, \sigma^2) : \sigma \in (0, \infty)\}$$

- Here $\Lambda = (0, \infty)$.
- Data $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} N(0, (\sigma^*)^2) \in \mathcal{N}$.

Example: parametric model

$$\mathcal{N} = \{N(0, \sigma^2) : \sigma \in (0, \infty)\}$$

- Here $\Lambda = (0, \infty)$.
- Data $X = (X_1, \ldots, X_n) \stackrel{\text{IID}}{\sim} N(0, (\sigma^*)^2) \in \mathcal{N}$.
- The parameter map is just $\theta(\sigma) = \sigma^2$.

Example: parametric model

$$\mathcal{N} = \{N(0, \sigma^2) : \sigma \in (0, \infty)\}$$

- Here $\Lambda = (0, \infty)$.
- Data $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} N(0, (\sigma^*)^2) \in \mathcal{N}$.
- The parameter map is just $\theta(\sigma) = \sigma^2$.
- An example statistic in this case is

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

Example: parametric model

$$\mathcal{N} = \{N(0, \sigma^2) : \sigma \in (0, \infty)\}$$

- Here $\Lambda = (0, \infty)$.
- Data $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} N(0, (\sigma^*)^2) \in \mathcal{N}$.
- The parameter map is just $\theta(\sigma) = \sigma^2$.
- An example statistic in this case is

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

- We say that $\hat{\Theta}$ is a point-estimator of $(\sigma^*)^2$.

Continuation

The bias is

$$\begin{aligned}\text{bias}(\hat{\Theta}) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] - (\sigma^*)^2 \\ &= -\frac{1}{n}(\sigma^*)^2\end{aligned}$$

Continuation

The bias is

$$\begin{aligned}\text{bias}(\hat{\Theta}) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] - (\sigma^*)^2 \\ &= -\frac{1}{n}(\sigma^*)^2\end{aligned}$$

- We call a point estimator **biased** if the bias is not zero.

Continuation

The bias is

$$\begin{aligned}\text{bias}(\hat{\Theta}) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] - (\sigma^*)^2 \\ &= -\frac{1}{n}(\sigma^*)^2\end{aligned}$$

- We call a point estimator **biased** if the bias is not zero.
- If the bias is zero we call it **unbiased**.

Continuation

The bias is

$$\begin{aligned}\mathrm{bias}(\hat{\Theta}) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] - (\sigma^*)^2 \\ &= -\frac{1}{n}(\sigma^*)^2\end{aligned}$$

- We call a point estimator **biased** if the bias is not zero.
- If the bias is zero we call it **unbiased**.
- If $\lim_{n\to\infty} \mathrm{bias}(\hat{\Theta}) = 0$ we say that the estimator is **asymptotically unbiased**.

Continuation

The bias is

$$\begin{aligned}\mathrm{bias}(\hat{\Theta}) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] - (\sigma^*)^2 \\ &= -\frac{1}{n}(\sigma^*)^2\end{aligned}$$

- We call a point estimator **biased** if the bias is not zero.
- If the bias is zero we call it **unbiased**.
- If $\lim_{n\to\infty} \mathrm{bias}(\hat{\Theta}) = 0$ we say that the estimator is **asymptotically unbiased**.
- If we change the estimator to

$$\hat{\Theta}_1(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

it becomes unbiased.

Example: non parametric model

$$\mathcal{E} = \{F(x;\lambda) : \mathbb{X} \to [0,1] : \lambda \in \Lambda,\ F \text{ is a DF}\}$$

- Consider the parameter map $\theta(\lambda) = \int x dF(x;\lambda)$, i.e. the expectation.
- Data $X = (X_1, \dots, X_n) \overset{\text{IID}}{\sim} F(\cdot;\lambda^*) \in \mathcal{E}$.
- An example statistic in this case is

$$\hat{\Theta}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- we say that $\hat{\Theta}$ is a point estimator of $\theta^* = \int x dF(\cdot;\lambda^*)$.

Again the bias is defined as

$$\text{bias}(\widehat{\Theta}(X)) := \mathbb{E}(\widehat{\Theta}(X)) - \theta^* = 0$$

Thus our estimator is unbiased!

Biased or Unbiased

Traditionally statistics has cared a lot about unbiased estimators, as they will give the correct value in average.

Question

Do you think that having a biased estimator could be better than having an unbiased estimator?

Terminology

Definition (Standard Error of a Point Estimator)

The standard deviation of the point estimator $\widehat{\Theta}_n(X)$ of $\theta^* \in \mathbf{\Theta}$ is called the **standard error**:

$$\mathsf{se}(\widehat{\Theta}_n(X)) := \sqrt{\mathbb{V}_{\lambda^*}(\widehat{\Theta}_n)} := \sqrt{\int \left(\widehat{\Theta}_n(x) - \mathbb{E}_{\lambda^*}(\widehat{\Theta}_n)\right)^2 dF(x;\lambda^*)} \ . \tag{2}$$

Bias and variance decomposition

Definition (Mean Squared Error (MSE) of a Point Estimator)

Often, the quality of a point estimator $\widehat{\Theta}$ of $\theta^* \in \mathbf{\Theta}$ is assessed by the **mean squared error** or MSE defined by:

$$\mathsf{MSE}(\widehat{\Theta}(X)) := \mathbb{E}_{\lambda^*}\left((\widehat{\Theta}(X) - \theta^*)^2\right) .$$

$$\mathsf{MSE}(\widehat{\Theta}) = (\mathsf{se}(\widehat{\Theta}))^2 + (\mathsf{bias}(\widehat{\Theta}))^2 . \tag{3}$$

Convergence of random variables

Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. We say that X_n converges to m **almost surely** if

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = m\right) = 1,$$

denoted as

$$X_n \stackrel{a.s.}{\to} m$$

Strong law of large numbers

Theorem (Strong law of large numbers)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let* $X_1, X_2, \ldots, \in L^2(\mathbb{P})$ *be a sequence of i.i.d. RVs with* $\mathbb{E}[X_i] = \mu$. *Then*

$$\overline{X}_n \overset{a.s.}{\to} \mu.$$

Convergence in probability

Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. We say that X_n converges to X in probability, and write:

$$X_n \xrightarrow{\mathbb{P}} X$$

if for every real number $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Asymptotic consistency

Definition (Asymptotic Consistency of a Point Estimator)

A point estimator $\widehat{\Theta}_n$ of $\theta^* \in \Theta$ is said to be **asymptotically consistent** if:

$$\widehat{\Theta}_n \xrightarrow{\mathbb{P}} \theta^*, \quad n \to \infty$$