Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recall from last time

- Concentration of measure is a statement of the form, for every $0 < \delta < 1$ there is an $\epsilon > 0$ such that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

Recall from last time

- Concentration of measure is a statement of the form, for every $0 < \delta < 1$ there is an $\epsilon > 0$ such that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

- Chebyschev, we only know variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{n\epsilon^2}$$

Recall from last time

- Concentration of measure is a statement of the form, for every $0 < \delta < 1$ there is an $\epsilon > 0$ such that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

- Chebyschev, we only know variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{n\epsilon^2}$$

- Hoeffding, we only know boundedness $a \leq X \leq b$

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Recall from last time

- Bennett, we know boundedness and variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2\exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.

Recall from last time

- Bennett, we know boundedness and variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2\exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

  where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.

- A confidence interval is a random interval I that is determined from $X_1, \ldots, X_n$ and satisfies

$$\mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta$$

Recall from last time

- Bennett, we know boundedness and variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2\exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

  where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.

- A confidence interval is a random interval I that is determined from $X_1, \ldots, X_n$ and satisfies

$$\mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta$$

- The confidence $1 - \delta$ tells us that **before** we compute the interval, the probability that our interval I covers $\mathbb{E}[X]$ is at least $1 - \delta$.

Risk

Model of the problem

Whatever assumptions we make about our experiment, is covered by the concept of a **statistical model**. I.e. it is what we assume is the truth

Risk

Model of the problem

Whatever assumptions we make about our experiment, is covered by the concept of a **statistical model**. I.e. it is what we assume is the truth

Definition

A **statistical model** is an indexed family of distributions (or densities or regression functions) $\mathcal{F} = \{f(x;\theta), \theta \in \Theta\}$.

Risk

Model of the problem

Whatever assumptions we make about our experiment, is covered by the concept of a **statistical model**. I.e. it is what we assume is the truth

Definition

A **statistical model** is an indexed family of distributions (or densities or regression functions) $\mathcal{F} = \{f(x;\theta), \theta \in \Theta\}$.

- A **parametric model** is a model where the indexing parameter θ is a vector in k-dimensional Euclidean space. That is, θ is finite dimensional.
- A **non-parametric model** is a model where Θ is infinite dimensional.

Risk

Example

$\mathcal{N} = \{N(\mu, \sigma), \mu \in \mathbb{R}, \sigma > 0\}$. This is a parametric model.

Risk

Example

$\mathcal{N} = \{N(\mu, \sigma), \mu \in \mathbb{R}, \sigma > 0\}$. This is a parametric model.

Example

$\mathcal{E} = \{F : F \text{ is a CDF}\}$. This is a non-parametric model.

Risk

Example

$\mathcal{N} = \{N(\mu, \sigma), \mu \in \mathbb{R}, \sigma > 0\}$. This is a parametric model.

Example

$\mathcal{E} = \{F : F \text{ is a CDF}\}$. This is a non-parametric model.

Example

$\mathcal{E} = \{F : F \text{ is a CDF and for } X \sim F, \mathbb{V}[X] < \infty\}$. This is a non-parametric model.

Risk

Example

$\mathcal{N} = \{N(\mu, \sigma), \mu \in \mathbb{R}, \sigma > 0\}$. This is a parametric model.

Example

$\mathcal{E} = \{F : F \text{ is a CDF}\}$. This is a non-parametric model.

Example

$\mathcal{E} = \{F : F \text{ is a CDF and for } X \sim F, \mathbb{V}[X] < \infty\}$. This is a non-parametric model.

Example

$\mathcal{E} = \{F : F \text{ is a CDF and for } X \sim F, \mathbb{P}[X \in (a, b)] = 1\}$. This is a non-parametric model.

Inference

Inference

Inference lies at the heart of statistics, learning and AI. The question is: what do we want to know?

Inference

Inference

Inference lies at the heart of statistics, learning and AI. The question is: what do we want to know?

The truth is out there

Under a statistical model $\mathcal{F}$, there is a hidden $f^* \in \mathcal{F}$ that generates the data, we would like to infer something about f^* using observations.

Here are some examples of inference problems:

Estimation of the distribution function

This is the fundamental problem. Once we have the distribution function we have everything!

Often we can only get it up to some error, i.e. we get something like

$$\mathbb{P}(\hat{F} - \epsilon \leq F \leq \hat{F} + \epsilon) \geq 1 - \delta$$

This means it will not be good at estimating the density...

Here are some examples of inference problems:

Estimation of the distribution function

This is the fundamental problem. Once we have the distribution function we have everything!
Often we can only get it up to some error, i.e. we get something like

$$\mathbb{P}(\hat{F} - \epsilon \leq F \leq \hat{F} + \epsilon) \geq 1 - \delta$$

This means it will not be good at estimating the density...

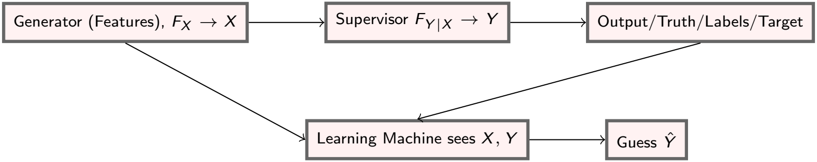
Estimating the density directly

If we can estimate the density up to some error, then we can also estimate the distribution function up to some error.
This is the **holy grail** of estimation.

Supervised learning

Setup

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



Generator (Features), $F_X \to X$
Supervisor $F_{Y|X} \to Y$
Output/Truth/Labels/Target
Learning Machine sees X, Y
Guess $\hat{Y}$

Examples

Example

Housing prices?

Examples

Example

Housing prices?

- G, the proceduce that selects a house and gives some data about the house

Examples

Example

Housing prices?

- G, the proceduce that selects a house and gives some data about the house
- S, the process by which given the house produces the price

Examples

Example

Housing prices?

- G, the proceduce that selects a house and gives some data about the house
- S, the process by which given the house produces the price
- LM, The Learning Machine observing pairs of data of house and price.

Examples

Example

SMS spam and not spam

- G, the recieving of a sms

Examples

Example

SMS spam and not spam

- G, the recieving of a sms
- S, You saying if the text is spam or not

Examples

Example

SMS spam and not spam

- G, the recieving of a sms
- S, You saying if the text is spam or not
- LM, The Learning Machine observing pairs of text and spam/ham. We "learned" that free appearing in the text is a good predictor, so we could use that as our guess: free → guess spam.

What is not supervised

If there is no supervisor there is no supervision

Lets say that all we have access to is the output of the generator. For instance house data. This is not a supervised problem as there is no supervisor, **no labels**.

??

Which of the following is supervised learning?

- You have a bunch of photos of 6 people but without information about who is on which one and you want to divide this dataset into 6 piles, each with the photos of one individual.

??

Which of the following is supervised learning?

- You have a bunch of photos of 6 people but without information about who is on which one and you want to divide this dataset into 6 piles, each with the photos of one individual.
- You get a bunch of photos with information about what is on them and then you train a model to recognize new photos.

??

Which of the following is supervised learning?

- You have a bunch of photos of 6 people but without information about who is on which one and you want to divide this dataset into 6 piles, each with the photos of one individual.
- You get a bunch of photos with information about what is on them and then you train a model to recognize new photos.
- You have a bunch of molecules and information about which are drugs and you train a model to answer whether a new molecule is also a drug.

??

Which of the following is supervised learning?

- You have a bunch of photos of 6 people but without information about who is on which one and you want to divide this dataset into 6 piles, each with the photos of one individual.
- You get a bunch of photos with information about what is on them and then you train a model to recognize new photos.
- You have a bunch of molecules and information about which are drugs and you train a model to answer whether a new molecule is also a drug.
- You have molecules, part of them are drugs and part are not but you do not know which are which and you want the algorithm to discover the drugs.

Mathematical formulation

- Generator: represented by F_X.
- Supervisor: represented by $F_{Y|X}$
- Learning Machine: Trying to learn the relation between X, Y and use that to guess Y given X.

Risk and loss functions

Loss

Much of machine learning is centered around a loss function. A loss function measures quality of an estimate (the Learning Machine).

Risk and loss functions

Loss

Much of machine learning is centered around a loss function. A loss function measures quality of an estimate (the Learning Machine).

Let g the the guessing function of the LM, i.e. it takes X and spits out a guess for Y, a loss function takes a point and our function and spits out a value $L((x, y), g)$

Risk and loss functions

Loss

Much of machine learning is centered around a loss function. A loss function measures quality of an estimate (the Learning Machine).

Let g the the guessing function of the LM, i.e. it takes X and spits out a guess for Y, a loss function takes a point and our function and spits out a value $L((x,y),g)$

Example

Quadratic loss

$$L((x,y),g) = (y - g(x))^2$$

Risk and loss functions

Loss

Much of machine learning is centered around a loss function. A loss function measures quality of an estimate (the Learning Machine).

Let g the the guessing function of the LM, i.e. it takes X and spits out a guess for Y, a loss function takes a point and our function and spits out a value $L((x,y),g)$

Example

Quadratic loss

$$L((x,y),g) = (y - g(x))^2$$

Risk

The risk is expected loss

$$R(g) = \mathbb{E}[L((X,Y),g)]$$

Risk and loss functions

Example

In the case of quadratic loss we simply get

$$R(g) = \mathbb{E}[(Y - g(X))^2]$$

Risk and loss functions

Example

In the case of quadratic loss we simply get

$$R(g) = \mathbb{E}[(Y - g(X))^2]$$

Goal of the LM

The goal of the Learning machine is to minimize its risk!
In the quadratic case this is called least squares. Using what?

Different problems

Find f

In the problem "find f", the supervisor uses f to construct Y, i.e. $Y = f(X)$ and the learning machine tries to find f.

Different problems

Find f

In the problem "find f", the supervisor uses f to construct Y, i.e. $Y = f(X)$ and the learning machine tries to find f.

Regression

In the regression problem, the supervisor has $F_{Y|X}$ but the Learning machine only tries to learn $r(X) = \mathbb{E}[Y \mid X]$.

Different problems

Find f

In the problem "find f", the supervisor uses f to construct Y, i.e. $Y = f(X)$ and the learning machine tries to find f.

Regression

In the regression problem, the supervisor has $F_{Y|X}$ but the Learning machine only tries to learn $r(X) = \mathbb{E}[Y \mid X]$.

Pattern recognition

The supervisor uses a discrete distribution for Y, i.e. $F_{Y|X}$ is a discrete distribution. Here the learning machine uses $0 - 1$ loss.

Regression

Regression function

$$r(x) = \int y dF_{Y|X}(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

Regression

Regression function

$$r(x) = \int y dF_{Y|X}(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

We need some things to be defined

- We should specify the statistical model $\mathcal{F}$ to say our assumptions.

Regression

Regression function

$$r(x) = \int y dF_{Y|X}(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

We need some things to be defined

- We should specify the statistical model $\mathcal{F}$ to say our assumptions.
- We should specify a loss function L. Quadratic loss most common.

Regression

Regression function

$$r(x) = \int y dF_{Y|X}(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

We need some things to be defined

- We should specify the statistical model $\mathcal{F}$ to say our assumptions.
- We should specify a loss function L. Quadratic loss most common.
- We should specify a so called "model space", where are we searching? Linear functions? We denote this by $\mathcal{M}$.

The risk minimization problem

The learning machine wants to solve the following problem

$$g^* = \arg\min_{g \in \mathcal{M}} R(g)$$

Goal

The learning machine tries to minimize the risk among the possible models in $\mathcal{M}$. In linear regression we where searching among functions of the type $g_{(k,m)}(x) = kx + m$.