Recitation 2

Statistical Learning Theory

Artie Shen and Brett Bernstein

CDS

February 5, 2020

Motivation

In data science problems, we generally need to:

- Make a prediction
- Take an action
- Produce an outcome
- Evaluate the "quality" of the prediction / action

But how can we formalize this?

Formalization

The Spaces

$\mathcal{X}$: input space $\mathcal{Y}$: outcome space $\mathcal{A}$: action space

Prediction Function

A **prediction function** f gets an input $x \in \mathcal{X}$ and produces an action $a \in \mathcal{A}$:

$$f : \mathcal{X} \mapsto \mathcal{A}$$

Loss Function

A **loss function** $\ell(a, y)$ evaluates an action $a \in \mathcal{A}$ in the context of an outcome $y \in \mathcal{Y}$:

$$\ell : \mathcal{A} \times \mathcal{Y} \mapsto \mathbb{R}$$

Risk Function

- Given a loss function ℓ, how can we evaluate the "average performance" of a prediction function f?
- To do so, we need to first assume that there is a **data generating distribution** $\mathcal{P}_{x,y}$.
- Then the expected loss of f on $\mathcal{P}_{x,y}$ will relect the notion of "average preformance".

Definition

The **risk** of a prediction function $f : \mathcal{X} \mapsto \mathcal{A}$ is

$$R(f) = \mathbb{E}\ell(f(x), y)$$

It is the expected loss of f on a new sample (x, y) drawn from $\mathcal{P}_{X,Y}$.

The Bayes Prediction Function

Definition

A **Bayes prediction function** $f^* : \mathcal{X} \mapsto \mathcal{Y}$ is a function that achieves the *minimal risk* among all possible functions:

$$f^* \in \arg\min_f R(f),$$

where the minimum is taken from all functions that maps from $\mathcal{X}$ to $\mathcal{A}$.

The risk of a Bayes function is called **Bayes risk**.

Example: Least Square Regression

- Spaces: $\mathcal{A} = \mathcal{Y} = \mathbb{R}$
- Loss function:

$$\ell(a, y) = (a - y)^2$$

- Risk:

$$\begin{aligned} R(f) &= \mathbb{E}[(f(x) - y)^2] \\ (homework) &= \mathbb{E}[(f(x) - \mathbb{E}[y|x])^2] + \mathbb{E}[(y - \mathbb{E}[y|x])^2] \end{aligned}$$

- So the Bayes function is

$$f^*(x) = \mathbb{E}[y|x]$$

Example: Multiclass Classification

- Spaces: $\mathcal{A} = \mathcal{Y} = \{1, ..., k\}$
- Loss function:

$$\ell(a, y) = 1(a \neq y) := \begin{cases} 1 & \text{if } a \neq y \\ 0 & \text{otherwise} \end{cases}$$

- Risk:

$$\begin{aligned} R(f) &= \mathbb{E}[1(f(x) \neq y)] \\ &= 0 \cdot P(f(x) = y) + 1 \cdot P(f(x) \neq y) \end{aligned}$$

- The Bayes function is just the assigment to the most likely class

$$f^*(x) \in \underset{1 \leq c \leq k}{\arg\max}\, P(y = c|x)$$

Case Study

Data Generating Distribution

We are given a generative process:

$$y = ax^2 + bx + c$$

where $a \sim \mathcal{N}(\mu_a, \sigma_a^2)$, $b \sim \mathcal{N}(\mu_b, \sigma_b^2)$, $c \sim \mathcal{N}(\mu_c, \sigma_c^2)$, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$. For the purposes of this lab, let's say $\mu_a = 1, \mu_b = 2, \mu_c = 3, \mu_x = 0$ and $\sigma_a = \sigma_b = \sigma_c = \sigma_x = 1$.

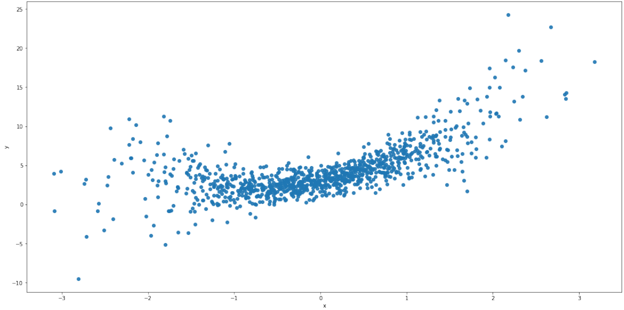
Case Study

Data Generating Distribution

We are given a generative process:

$$y = ax^2 + bx + c$$

where $a \sim \mathcal{N}(\mu_a, \sigma_a^2)$, $b \sim \mathcal{N}(\mu_b, \sigma_b^2)$, $c \sim \mathcal{N}(\mu_c, \sigma_c^2)$, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$.



Case Study

Data Generating Distribution

We are given a generative process:

$$y = ax^2 + bx + c$$

where $a \sim \mathcal{N}(\mu_a, \sigma_a^2)$, $b \sim \mathcal{N}(\mu_b, \sigma_b^2)$, $c \sim \mathcal{N}(\mu_c, \sigma_c^2)$, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Suppose the input, action, and output space are $\mathbb{R}$.

- What is an appropriate loss function to parameterize the risk $R(f)$?
- Can we find the Bayes prediction function $f^*(x) : \mathbb{R} \mapsto \mathbb{R}$
- What is the Bayes risk $R(f^*(x))$ associated with the Bayes prediction function?

The Risk Function

Recap

The **risk** of a prediction function $f : \mathcal{X} \mapsto \mathcal{A}$ is

$$R(f) = \mathbb{E}\ell(f(x), y)$$

It is the expected loss of f on a new sample (x, y) drawn from $\mathcal{P}_{X,Y}$.

- Since it is a regression problem, let's use the ℓ_2 loss $\ell(f(x), y) = (f(x) - y)^2$.
- The risk can be than expressed as $R(f) = \mathbb{E}[(f(x) - \mathbb{E}[y|x])^2] + \mathbb{E}[(y - E[y|x])^2]$.

The Bayes Prediction Function

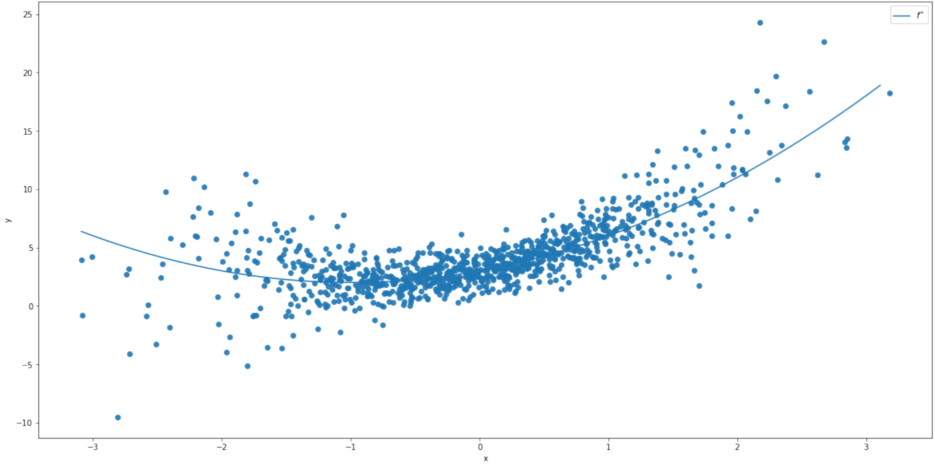
Recap

A **Bayes prediction function** f^* is a function that achieves the minimal risk among all possible functions:

$$f^* \in \arg\min_f R(f)$$

- With ℓ_2 loss, the Bayes prediction function is $f^*(x) = \mathbb{E}[y|x] = \mu_a x^2 + \mu_b x + \mu_c$.
- Note that $f^*(x)$ is independent of the distribution of X.

$$f^*(x) = \mathbb{E}[y|x] = \mu_a x^2 + \mu_b x + \mu_c$$



Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$R(f^*(x)) = \mathbb{E}_{x,y}[(y - \mathbb{E}[y|x])^2]$$

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$
\begin{aligned}
R(f^*(x)) &= \mathbb{E}_{x,y}[(y - \mathbb{E}[y|x])^2] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}[(y - \mathbb{E}[y|x])^2]] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}(ax^2 + bx + c - \mu_a x^2 - \mu_b x - \mu_c)^2] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}((a - \mu_a)x^2 + (b - \mu_b)x + (c - \mu_c))^2]
\end{aligned}
$$

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$
\begin{aligned}
R(f^*(x)) &= \mathbb{E}_{x,y}[(y - \mathbb{E}[y|x])^2] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}[(y - \mathbb{E}[y|x])^2]] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}(ax^2 + bx + c - \mu_a x^2 - \mu_b x - \mu_c)^2] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}((a - \mu_a)x^2 + (b - \mu_b)x + (c - \mu_c))^2] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}[A^2x^4 + 2ABx^3 + (AB + AC)x^2 + 2BCx + 2C^2]
\end{aligned}
$$

For simplicity, let's say $A = a - \mu_a$, $B = b - \mu_b$, and $C = c - \mu_c$.

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

Remember that $x \sim \mathcal{N}(0, 1)$, then:

$$\mathbb{E}[x] = \mu_x = 0$$

$$\mathbb{E}[x^2] = \mu_x^2 + \sigma_x^2 = 1$$

$$\mathbb{E}[x^3] = \mu_x(\mu_x^2 + 3\sigma_x^2) = 0$$

$$\mathbb{E}[x^4] = \mu_x^4 + 6\mu_x^2\sigma_x^2 + 3\sigma_x^2 = 3$$

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$R(f^*(x)) = \mathbb{E}_x[\mathbb{E}_{y|x}[A^2x^4 + 2ABx^3 + (AB + AC)x^2 + 2BCx + 2C^2]$$

For simplicity, let's say $A = a - \mu_a$, $B = b - \mu_b$, and $C = c - \mu_c$.

- A, B, C are independent of x and therefore $\mathbb{E}_x\mathbb{E}_{x|y} f(A, B, C)x = \mathbb{E}_x[x] \cdot \mathbb{E}_{y|x}[f(A, B, C)]$

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$R(f^*(x)) = \mathbb{E}_x[\mathbb{E}_{y|x}[A^2x^4 + 2ABx^3 + (AB + AC)x^2 + 2BCx + 2C^2]$$

For simplicity, let's say $A = a - \mu_a$, $B = b - \mu_b$, and $C = c - \mu_c$.

- A, B, C are independent of x and therefore $\mathbb{E}_x\mathbb{E}_{x|y}f(A, B, C)x = \mathbb{E}_x[x] \cdot \mathbb{E}_{y|x}[f(A, B, C)]$
- $\mathbb{E}_x\mathbb{E}_{x|y}[2ABx^3] = \mathbb{E}_x[x^3] \cdot \mathbb{E}_{y|x}[2AB] = 0$ and $\mathbb{E}_x\mathbb{E}_{x|y}[2BCx] = 0$

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$R(f^*(x)) = \mathbb{E}_x[\mathbb{E}_{y|x}[A^2x^4 + 2ABx^3 + (AB + AC)x^2 + 2BCx + 2C^2]$$

For simplicity, let's say $A = a - \mu_a$, $B = b - \mu_b$, and $C = c - \mu_c$.

- A, B, C are independent of x and therefore $\mathbb{E}_x\mathbb{E}_{x|y}f(A, B, C)x = \mathbb{E}_x[x] \cdot \mathbb{E}_{y|x}[f(A, B, C)]$
- $\mathbb{E}_x\mathbb{E}_{x|y}[2ABx^3] = \mathbb{E}_x[x^3] \cdot \mathbb{E}_{y|x}[2AB] = 0$ and $\mathbb{E}_x\mathbb{E}_{x|y}[2BCx] = 0$
- $\mathbb{E}_{y|x}[AB] = \mathbb{E}[(a - \mu_a)(b - \mu_b)] = Cov(a, b) = 0$ because a and b are independently draw from Gaussian.

Bayes Risk

Question

What is the Bayes risk $R(f^*(x))$?

$$
\begin{aligned}
R(f^*(x)) &= \mathbb{E}_x[\mathbb{E}_{y|x}[A^2x^4 + 2ABx^3 + (AB+AC)x^2 + 2BCx + 2C^2]] \\
&= \mathbb{E}_x[\mathbb{E}_{y|x}[A^2x^4 + 2C^2]] \\
&= \mathbb{E}_{y|x}[A^2] \cdot \mathbb{E}_x[x^4] + 2 \cdot \mathbb{E}_{y|x}[C^2] \\
&= \mathbb{E}_{y|x}[(a-\mu_a)^2] \cdot 3 + 2 \cdot \mathbb{E}_{y|x}[(c-\mu_c)^2] \\
&= \sigma_a^2 \cdot 3 + 2 \cdot \sigma_c^2 \\
&= 5
\end{aligned}
$$

Empirical Risk

Recap

The **empirical risk** of f with respect to a dataset $\mathcal{D}$ is

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i), y_i)$$

- In reality, we never know $\mathcal{P}_{x,y}$ and we work with finite samples.
- Usually, we build our predictive model by finding the best predictor $\hat{f}(x)$ that minimizes the empirical risk.
- We will study the risk and empirical risk of two simple prediction functions.

Predict by Memorizing

Method 1

One way to achieve $\hat{R}_n(f) = 0$ on a dataset $\mathcal{D}_n$ is to memorize the data:

$$\hat{f}_m(x) = \begin{cases} y_i & x \in \mathcal{D}_n \\ 0 & otherwise \end{cases}$$

- What's its empirical risk $\hat{R}_n(\hat{f}_m)$ on $\mathcal{D}_n$?
- What's the risk $R(\hat{f}_m)$?
- What's $\mathbb{E}[\hat{f}_m(x)]$?

Predict by Memorizing

Derive the risk $R_n(f_m(x))$ of the predictor $\hat{f}_m(x)$ that memorizes the data:

$$\begin{aligned} R_n(\hat{f}_m(x)) &= \mathbb{E}_{x,y}[(y - \hat{f}_m(x))^2] \\ &= \mathbb{E}_{x,y}[(y - 0)^2] \\ &= \mathbb{E}_x \mathbb{E}_{y|x}[y^2] \\ &= \mathrm{VAR}(y) + \mathbb{E}[y]^2 \end{aligned}$$

Predict by Memorizing

Derive the risk $R_n(\hat{f}_m(x))$ of the predictor $\hat{f}_m(x)$ that memorizes the data:

$$\begin{aligned}
\hat{R}_n(\hat{f}_m(x)) &= \mathbb{E}_{x,y}[(y - \hat{f}_m(x))^2] \\
&= \mathbb{E}_{x,y}[(y-0)^2] \\
&= \mathbb{E}_x \mathbb{E}_{y|x}[y^2] \\
&= \mathrm{VAR}(y) + \mathbb{E}[y]^2 \\
&= \mathbb{E}[\sigma_a^2 x^4 + \sigma_b^2 x^2 + \sigma_c^2 + (\mu_a x^2 + \mu_b x + \mu_c)^2] \\
&= \ldots \\
&= 3(\sigma_a^2 + \mu_a^2) + (\sigma_b^2 + \mu_b^2 + 2\mu_c\mu_a) + \sigma_c^2 + \mu_c^2 \\
&= 27
\end{aligned}$$

A Linear Model

Method 2

Now, let's consider a linear model:

$$\hat{f}_l(x) = \alpha x + \beta$$

- What's risk $R(\hat{f}_l(x))$ of $\hat{f}_l(x)$?
- What are the α^* and β^* that minimises $R(\hat{f}_l(x))$?

Risk of the linear model

$$
\begin{aligned}
R(\hat{f}_l) &= \mathbb{E}[(y - \alpha x - \beta)^2] \\
&= \mathbb{E}_x \mathbb{E}_{y|x}[(Y - \alpha x - \beta)^2] \\
&= \mathbb{E}_x[\mathrm{VAR}(Y - \alpha x - \beta) + \mathbb{E}^2_{y|x}[Y - \alpha x - \beta]] \\
&= \ldots \\
&= 3(\sigma_a^2 + \mu_a^2) + (\sigma_b^2 + (\mu_b - \alpha)^2 + 2(\mu_c - \beta)\mu_a) + \sigma_c^2 + (\mu_c - \beta)^2
\end{aligned}
$$

Find the α^* and β^*

To find α^*, we need to solve :

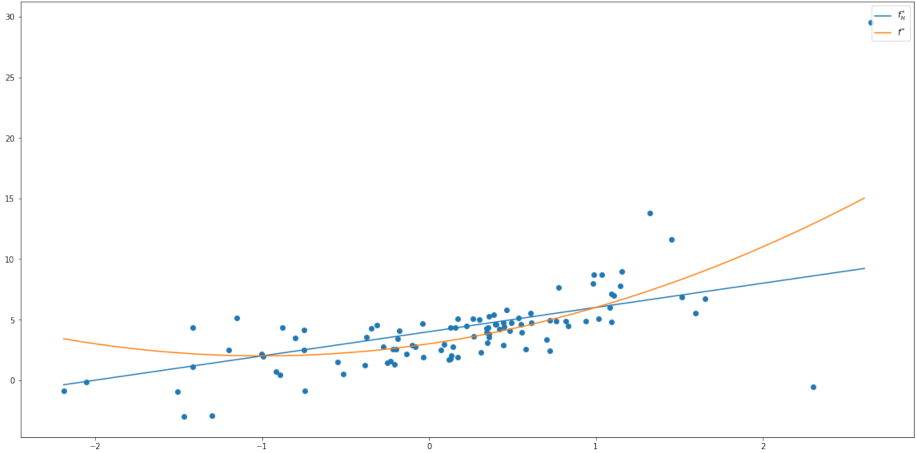
$$\frac{d}{d\alpha}R(f) = 2(\alpha - \mu_b) = 0$$

which gives us $\alpha^* = \mu_b$.
To find β^*, we need to solve :

$$\frac{d}{d\beta}R(f) = -2\mu_a + 2(\beta - \mu_c) = 0$$

which gives us $\beta^* = \mu_a + \mu_c$.

This is how $f_l^*(x) = \alpha^* x + \beta^*$ looks like (on a smaller set of data):



Coding Exercise

In the provided notebook, we will use Python to:

- create and sample a dataset $\mathcal{D}_n$ from the generative process
- calculate the empirical risk $\hat{R}_n(f^*(x))$ on $\mathcal{D}_n$ using $f^*(x)$ and compare it with the derived Bayesian risk
- create the model that memorizes $\mathcal{D}_n$ and calculate its empirical risk
- create the linear model and calculate its empirical risk