YSU Statistical ML, Fall 2019

Lecture 02

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

16 November 2019

Contents

- Conditional Distributions
- Bayes Predictor for the Classification Problem

Last Lecture Recap

Recall that our Learning Problem was stated as: We are given

- A Dataset of Observations $(\mathbf{x}_k, y_k)$, $k = 1, .., n$, coming as a realization of $(\mathbf{X}_k, Y_k)$ from an unknown Distribution $\mathcal{F}$;
- A Loss Function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$;
- A Predictive Model (set of Functions) $\mathcal{G}$

and we want to find $g^* \in \mathcal{G}$ such that

$$g^* \in \underset{g \in \mathcal{G}}{argmin}\, Risk(g) = \underset{g \in \mathcal{G}}{argmin}\, \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big).$$

Last Lecture Recap

Recall that our Learning Problem was stated as: We are given

- A Dataset of Observations $(\mathbf{x}_k, y_k)$, $k = 1, .., n$, coming as a realization of $(\mathbf{X}_k, Y_k)$ from an unknown Distribution $\mathcal{F}$;
- A Loss Function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$;
- A Predictive Model (set of Functions) $\mathcal{G}$

and we want to find $g^* \in \mathcal{G}$ such that

$$g^* \in \underset{g \in \mathcal{G}}{argmin}\, Risk(g) = \underset{g \in \mathcal{G}}{argmin}\, \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big).$$

And if g^* is the solution to this minimization problem for $\mathcal{G} = \{\text{all meausrable functions}\}$, then we will call g^* the **Bayes Predictor**.

Last Lecture Recap

Recall that our Learning Problem was stated as: We are given

- A Dataset of Observations $(\mathbf{x}_k, y_k)$, $k = 1, .., n$, coming as a realization of $(\mathbf{X}_k, Y_k)$ from an unknown Distribution $\mathcal{F}$;
- A Loss Function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$;
- A Predictive Model (set of Functions) $\mathcal{G}$

and we want to find $g^* \in \mathcal{G}$ such that

$$g^* \in \underset{g \in \mathcal{G}}{argmin}\, Risk(g) = \underset{g \in \mathcal{G}}{argmin}\, \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big).$$

And if g^* is the solution to this minimization problem for $\mathcal{G} = \{$all meausrable functions$\}$, then we will call g^* the **Bayes Predictor**. And if g^* is a Bayes Predictor, then its Risk, $Risk(g^*)$, is called the **Bayes Risk**.

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Now, if g is any Predictor (any function $g : \mathcal{X} \to \mathcal{Y}$), then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}(Y \neq g(\mathbf{X}))\Big) =$$

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Now, if g is any Predictor (any function $g : \mathcal{X} \to \mathcal{Y}$), then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}(Y \neq g(\mathbf{X}))\Big) = \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Now, if g is any Predictor (any function $g : \mathcal{X} \to \mathcal{Y}$), then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}(Y \neq g(\mathbf{X}))\Big) = \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

So the Risk of g in the case of Binary Classification with $0 - 1$ Loss is **the Probability to predict incorrectly**.

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Now, if g is any Predictor (any function $g : \mathcal{X} \to \mathcal{Y}$), then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}(Y \neq g(\mathbf{X}))\Big) = \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

So the Risk of g in the case of Binary Classification with $0 - 1$ Loss is **the Probability to predict incorrectly**.

And the Problem to find the Bayes Predictor is to find a function g with minimal Probability of incorrect Prediction:

$$g^* \in \underset{g}{argmin}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Now, if g is any Predictor (any function $g : \mathcal{X} \to \mathcal{Y}$), then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}(Y \neq g(\mathbf{X}))\Big) = \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

So the Risk of g in the case of Binary Classification with $0 - 1$ Loss is **the Probability to predict incorrectly**.

And the Problem to find the Bayes Predictor is to find a function g with minimal Probability of incorrect Prediction:

$$g^* \in \underset{g}{argmin}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Can you guess g^* ?

Example: Binary Classification Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$);
- our Loss Function is the $0 - 1$ loss: $\ell(y_1, y_2) = \mathbf{1}(y_1 \neq y_2)$

Now, if g is any Predictor (any function $g : \mathcal{X} \to \mathcal{Y}$), then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big(\mathbf{1}(Y \neq g(\mathbf{X}))\Big) = \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

So the Risk of g in the case of Binary Classification with $0 - 1$ Loss is **the Probability to predict incorrectly**.

And the Problem to find the Bayes Predictor is to find a function g with minimal Probability of incorrect Prediction:

$$g^* \in \underset{g}{argmin}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Can you guess g^* ? We will find g^* soon $\ddot\smile$.

Example: Least Squares Regression Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} \subset \mathbb{R}$;

Example: Least Squares Regression Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} \subset \mathbb{R}$;
- our Loss Function is the Quadratic Loss: $\ell(y_1, y_2) = (y_1 - y_2)^2$

Example: Least Squares Regression Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} \subset \mathbb{R}$;
- our Loss Function is the Quadratic Loss: $\ell(y_1, y_2) = (y_1 - y_2)^2$

And, if g is any Predictor, then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big((Y - g(\mathbf{X}))^2\Big).$$

Example: Least Squares Regression Problem

In this case we have:

- $\mathcal{X}$ is arbitrary, $\mathcal{Y} \subset \mathbb{R}$;
- our Loss Function is the Quadratic Loss: $\ell(y_1, y_2) = (y_1 - y_2)^2$

And, if g is any Predictor, then

$$Risk(g) = \mathbb{E}\Big(\ell(Y, g(\mathbf{X}))\Big) = \mathbb{E}\Big((Y - g(\mathbf{X}))^2\Big).$$

The Problem to find the Bayes Predictor in this case is:

$$g^* \in \underset{g}{argmin}\, \mathbb{E}\Big((Y - g(\mathbf{X}))^2\Big).$$

Conditional Distribution, Discrete Case

To continue with finding Bayes Predictors, we need the notion of Conditional Distribution.

Conditional Distribution, Discrete Case

To continue with finding Bayes Predictors, we need the notion of Conditional Distribution.

First assume X and Y are Jointly Distributed Discrete r.v. with the PMF $f_{X,Y}(x, y)$, i.e.,

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

Conditional Distribution, Discrete Case

To continue with finding Bayes Predictors, we need the notion of Conditional Distribution.

First assume X and Y are Jointly Distributed Discrete r.v. with the PMF $f_{X,Y}(x, y)$, i.e.,

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

The Marginal PMFs for X and Y will be:

$$f_X(x) = \mathbb{P}(X = x) = \sum_y f_{X,Y}(x, y),$$

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x f_{X,Y}(x, y).$$

Conditional Distribution, Discrete Case

To continue with finding Bayes Predictors, we need the notion of Conditional Distribution.

First assume X and Y are Jointly Distributed Discrete r.v. with the PMF $f_{X,Y}(x, y)$, i.e.,

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

The Marginal PMFs for X and Y will be:

$$f_X(x) = \mathbb{P}(X = x) = \sum_y f_{X,Y}(x, y),$$

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x f_{X,Y}(x, y).$$

Next, we know that

$$\mathbb{P}(Y = y|X = x) =$$

Conditional Distribution, Discrete Case

To continue with finding Bayes Predictors, we need the notion of Conditional Distribution.

First assume X and Y are Jointly Distributed Discrete r.v. with the PMF $f_{X,Y}(x,y)$, i.e.,

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

The Marginal PMFs for X and Y will be:

$$f_X(x) = \mathbb{P}(X = x) = \sum_y f_{X,Y}(x,y),$$

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x f_{X,Y}(x,y).$$

Next, we know that

$$\mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)},$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Sometimes we write this as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Sometimes we write this as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Now, we can calculate, say

$$\mathbb{P}(Y \in A | X = x) =$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Sometimes we write this as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Now, we can calculate, say

$$\mathbb{P}(Y \in A | X = x) = \sum_{y \in A} f_{Y|X}(y|x) =$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Sometimes we write this as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Now, we can calculate, say

$$\mathbb{P}(Y \in A|X = x) = \sum_{y \in A} f_{Y|X}(y|x) = \sum_{y \in A} f_{Y|X=x}(y).$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Sometimes we write this as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Now, we can calculate, say

$$\mathbb{P}(Y \in A|X = x) = \sum_{y \in A} f_{Y|X}(y|x) = \sum_{y \in A} f_{Y|X=x}(y).$$

Also, we can calculate, say, the Expected value of Y given the value of X:

$$\mathbb{E}(Y|X = x) =$$

Conditional Distribution, Discrete Case

Hence, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\sum_y f_{X,Y}(x,y)}.$$

Sometimes we write this as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Now, we can calculate, say

$$\mathbb{P}(Y \in A|X = x) = \sum_{y \in A} f_{Y|X}(y|x) = \sum_{y \in A} f_{Y|X=x}(y).$$

Also, we can calculate, say, the Expected value of Y given the value of X:

$$\mathbb{E}(Y|X = x) = \sum_y y \cdot f_{Y|X}(y|x) = \sum_y y \cdot f_{Y|X=x}(y)$$

Example

Conditional Distribution, Continuous Case

Now, if X and Y are Jointly Distributed Continuous r.v. with the PDF $f_{X,Y}(x, y)$, i.e.,

$$\mathbb{P}((X, Y) \in A) = \iint_{(x,y)\in A} f_{X,Y}(x, y)dxdy.$$

Conditional Distribution, Continuous Case

Now, if X and Y are Jointly Distributed Continuous r.v. with the PDF $f_{X,Y}(x,y)$, i.e.,

$$\mathbb{P}((X,Y) \in A) = \iint_{(x,y)\in A} f_{X,Y}(x,y)dxdy.$$

Then the Marginal PMFs for X and Y will be:

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx.$$

Conditional Distribution, Continuous Case

Now, if X and Y are Jointly Distributed Continuous r.v. with the PDF $f_{X,Y}(x,y)$, i.e.,

$$\mathbb{P}((X,Y) \in A) = \iint_{(x,y)\in A} f_{X,Y}(x,y)dxdy.$$

Then the Marginal PMFs for X and Y will be:

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx.$$

Now, in the analogy of the Discrete case, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\int_{\mathbb{R}} f_{X,Y}(x,y)dy},$$

for all x such that $f_X(x) \neq 0$.

Conditional Distribution, Continuous Case

Now, if X and Y are Jointly Distributed Continuous r.v. with the PDF $f_{X,Y}(x,y)$, i.e.,

$$\mathbb{P}((X,Y) \in A) = \iint_{(x,y)\in A} f_{X,Y}(x,y)\,dxdy.$$

Then the Marginal PMFs for X and Y will be:

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx.$$

Now, in the analogy of the Discrete case, we define the Conditional PMF of Y given X by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\int_{\mathbb{R}} f_{X,Y}(x,y)dy},$$

for all x such that $f_X(x) \neq 0$. Again, we write this also as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Conditional Distribution, Continuous Case

Now, we can calculate, say

$$\mathbb{P}(Y \in A | X = x) =$$

Conditional Distribution, Continuous Case

Now, we can calculate, say

$$\mathbb{P}(Y \in A | X = x) = \int_A f_{Y|X}(y|x)\, dy =$$

Conditional Distribution, Continuous Case

Now, we can calculate, say

$$\mathbb{P}(Y \in A | X = x) = \int_A f_{Y|X}(y|x)\, dy = \int_A f_{Y|X=x}(y)\, dy.$$

Conditional Distribution, Continuous Case

Now, we can calculate, say

$$\mathbb{P}(Y \in A|X = x) = \int_A f_{Y|X}(y|x)\,dy = \int_A f_{Y|X=x}(y)\,dy.$$

And the Expected value of Y given the value of X will be:

$$\mathbb{E}(Y|X = x) =$$

Conditional Distribution, Continuous Case

Now, we can calculate, say

$$\mathbb{P}(Y \in A | X = x) = \int_A f_{Y|X}(y|x)\, dy = \int_A f_{Y|X=x}(y)\, dy.$$

And the Expected value of Y given the value of X will be:

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\, dy = \int_{\mathbb{R}} y \cdot f_{Y|X=x}(y)\, dy.$$

Example

Interpretation of the Conditional Distribution

Note that we can also define the Conditional Distribution in the case when one of X, Y is continuous and the other one is Discrete.

Interpretation of the Conditional Distribution

Note that we can also define the Conditional Distribution in the case when one of X, Y is continuous and the other one is Discrete.

Say, we can have X, Y on the same Probability Space with X being continuous and $Y \sim \textit{Bernoulli}(0.5)$.

Interpretation of the Conditional Distribution

Note that we can also define the Conditional Distribution in the case when one of X, Y is continuous and the other one is Discrete.

Say, we can have X, Y on the same Probability Space with X being continuous and $Y \sim Bernoulli(0.5)$. Then we can interpret $X|Y = y$ as the Distribution of X, given that $Y = y$, e.g., we can have

$$X|Y = 0 \sim \mathcal{N}(0, 1) \qquad \textit{and} \qquad X|Y = 1 \sim \mathcal{N}(2, 1)$$

Interpretation of the Conditional Distribution

Note that we can also define the Conditional Distribution in the case when one of X, Y is continuous and the other one is Discrete.

Say, we can have X, Y on the same Probability Space with X being continuous and $Y \sim Bernoulli(0.5)$. Then we can interpret $X|Y = y$ as the Distribution of X, given that $Y = y$, e.g., we can have

$$X|Y = 0 \sim \mathcal{N}(0, 1) \qquad \textit{and} \qquad X|Y = 1 \sim \mathcal{N}(2, 1)$$

We can think about this as: we have some values for X, coming from two classes, say, $Y = 0$ and $Y = 1$, i.e., with two Labels. The values of X, having the class 0 are $\mathcal{N}(0, 1)$, and for the class 1 they are $\mathcal{N}(2, 1)$.

Interpretation of the Conditional Distribution

Note that we can also define the Conditional Distribution in the case when one of X, Y is continuous and the other one is Discrete.

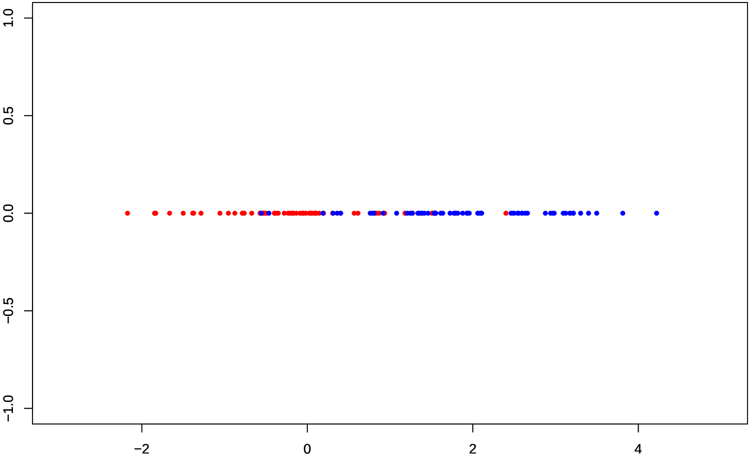
Say, we can have X, Y on the same Probability Space with X being continuous and $Y \sim \mathit{Bernoulli}(0.5)$. Then we can interpret $X|Y = y$ as the Distribution of X, given that $Y = y$, e.g., we can have

$$X|Y = 0 \sim \mathcal{N}(0, 1) \qquad \mathit{and} \qquad X|Y = 1 \sim \mathcal{N}(2, 1)$$

We can think about this as: we have some values for X, coming from two classes, say, $Y = 0$ and $Y = 1$, i.e., with two Labels. The values of X, having the class 0 are $\mathcal{N}(0, 1)$, and for the class 1 they are $\mathcal{N}(2, 1)$.

Here the Distribution of Y, in our case $\mathit{Bernoulli}(0.5)$ is called the Prior Class Distribution.

Example



Conditional Expectation as a RV

We have calculated above Conditional Expectations of the form

$$\mathbb{E}(X|Y=y).$$

Conditional Expectation as a RV

We have calculated above Conditional Expectations of the form

$$\mathbb{E}(X|Y = y).$$

Now, if we will not fix the value of Y, we will have a r.v. $\mathbb{E}(X|Y)$: this is something like "we are averaging over X for a given value of Y".

Conditional Expectation as a RV

We have calculated above Conditional Expectations of the form

$$\mathbb{E}(X|Y = y).$$

Now, if we will not fix the value of Y, we will have a r.v. $\mathbb{E}(X|Y)$: this is something like "we are averaging over X for a given value of Y". If we will denote it by $h(Y) = \mathbb{E}(X|Y)$, then for $Y = y$, this r.v. takes the value $h(y) = \mathbb{E}(X|Y = y)$.

Conditional Expectation as a RV

We have calculated above Conditional Expectations of the form

$$\mathbb{E}(X|Y = y).$$

Now, if we will not fix the value of Y, we will have a r.v. $\mathbb{E}(X|Y)$: this is something like "we are averaging over X for a given value of Y". If we will denote it by $h(Y) = \mathbb{E}(X|Y)$, then for $Y = y$, this r.v. takes the value $h(y) = \mathbb{E}(X|Y = y)$.

On important property of the Conditional Expectation is the following:

Theorem: If X and Y are r.v. with finite Expectation and Variance, then for any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\Big((X - \mathbb{E}(X|Y))^2\Big) \leq \mathbb{E}\Big((X - g(Y))^2\Big).$$

Conditional Expectation as a RV

We have calculated above Conditional Expectations of the form

$$\mathbb{E}(X|Y = y).$$

Now, if we will not fix the value of Y, we will have a r.v. $\mathbb{E}(X|Y)$: this is something like "we are averaging over X for a given value of Y". If we will denote it by $h(Y) = \mathbb{E}(X|Y)$, then for $Y = y$, this r.v. takes the value $h(y) = \mathbb{E}(X|Y = y)$.

On important property of the Conditional Expectation is the following:

Theorem: If X and Y are r.v. with finite Expectation and Variance, then for any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\Big((X - \mathbb{E}(X|Y))^2\Big) \leq \mathbb{E}\Big((X - g(Y))^2\Big).$$

In other words, if we have r.v.s X and Y, and we want to find a r.v. of the form $g(Y)$ which is the closest one to X in the MSE sense, then the best one is $\mathbb{E}(X|Y)$.

Bayes Predictor in the Binary Classification Problem

Let us go bact to our Binary Classification case with $\mathcal{Y} = \{0, 1\}$. Recall that, in the Binary Classification case, the the problem of finding the Bayes Predictor reduces to

$$g^* \in \underset{g}{argmin}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Bayes Predictor in the Binary Classification Problem

Let us go bact to our Binary Classification case with $\mathcal{Y} = \{0, 1\}$. Recall that, in the Binary Classification case, the the problem of finding the Bayes Predictor reduces to

$$g^* \in \underset{g}{\operatorname{argmin}}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Now we introduce the following function:

$$\eta(x) = \mathbb{E}(Y|X = x).$$

Bayes Predictor in the Binary Classification Problem

Let us go bact to our Binary Classification case with $\mathcal{Y} = \{0, 1\}$. Recall that, in the Binary Classification case, the the problem of finding the Bayes Predictor reduces to

$$g^* \in \underset{g}{argmin}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Now we introduce the following function:

$$\eta(x) = \mathbb{E}(Y|X = x).$$

Because Y is Binary, we will have

$\eta(x) = \mathbb{E}(Y|X = x) =$

Bayes Predictor in the Binary Classification Problem

Let us go bact to our Binary Classification case with $\mathcal{Y} = \{0, 1\}$. Recall that, in the Binary Classification case, the the problem of finding the Bayes Predictor reduces to

$$g^* \in \underset{g}{argmin}\, \mathbb{P}\Big(Y \neq g(\mathbf{X})\Big).$$

Now we introduce the following function:

$$\eta(x) = \mathbb{E}(Y|X = x).$$

Because Y is Binary, we will have

$$\eta(x) = \mathbb{E}(Y|X = x) = 0 \cdot \mathbb{P}(Y = 0|X = x) + 1 \cdot \mathbb{P}(Y = 1|X = x),$$

so

$$\eta(x) = \mathbb{P}(Y = 1|X = x).$$

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if } \ \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if } \ \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Then, g^* is a Bayes Predictor (Bayes Classifier).

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if} \quad \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Then, g^* is a Bayes Predictor (Bayes Classifier).

Proof: Assume g is any Predictor.

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if } \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Then, g^* is a Bayes Predictor (Bayes Classifier).

Proof: Assume g is any Predictor. Then we need to show that

$$Risk(g) \geq Risk(g^*),$$

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if} \quad \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Then, g^* is a Bayes Predictor (Bayes Classifier).

Proof: Assume g is any Predictor. Then we need to show that

$$Risk(g) \geq Risk(g^*),$$

i.e., that

$$\mathbb{P}(Y \neq g(X)) \geq \mathbb{P}(Y \neq g^*(X)),$$

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if } \ \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Then, g^* is a Bayes Predictor (Bayes Classifier).

Proof: Assume g is any Predictor. Then we need to show that

$$Risk(g) \geq Risk(g^*),$$

i.e., that

$$\mathbb{P}(Y \neq g(X)) \geq \mathbb{P}(Y \neq g^*(X)),$$

or, which is the same,

$$\mathbb{P}(Y = g(X)) \leq \mathbb{P}(Y = g^*(X)).$$

Bayes Predictor in the Binary Classification Problem

Now we consider the following Predictor:

$$g^*(x) = \begin{cases} 1, & \text{if} \quad \eta(x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Then, g^* is a Bayes Predictor (Bayes Classifier).

Proof: Assume g is any Predictor. Then we need to show that

$$Risk(g) \geq Risk(g^*),$$

i.e., that

$$\mathbb{P}(Y \neq g(X)) \geq \mathbb{P}(Y \neq g^*(X)),$$

or, which is the same,

$$\mathbb{P}(Y = g(X)) \leq \mathbb{P}(Y = g^*(X)).$$

So we need to find the maximum of $\mathbb{P}(Y = g(X))$ over all g.

Proof: Bayes Predictor in the BC Problem

Now, for a fixed x, consider

$$\mathbb{P}(Y = g(x) | X = x).$$

Proof: Bayes Predictor in the BC Problem

Now, for a fixed x, consider

$$\mathbb{P}(Y = g(x)|X = x).$$

We have

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \mathbb{P}(Y = 1|X = x), & \textit{if} \quad g(x) = 1 \\ \mathbb{P}(Y = 0|X = x), & \textit{if} \quad g(x) = 0 \end{cases}$$

Proof: Bayes Predictor in the BC Problem

Now, for a fixed x, consider

$$\mathbb{P}(Y = g(x)|X = x).$$

We have

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \mathbb{P}(Y = 1|X = x), & \textit{if} \quad g(x) = 1 \\ \mathbb{P}(Y = 0|X = x), & \textit{if} \quad g(x) = 0 \end{cases}$$

That is,

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \eta(x), & \textit{if} \quad g(x) = 1 \\ 1 - \eta(x), & \textit{if} \quad g(x) = 0 \end{cases}$$

Proof: Bayes Predictor in the BC Problem

Now, for a fixed x, consider

$$\mathbb{P}(Y = g(x)|X = x).$$

We have

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \mathbb{P}(Y = 1|X = x), & \textit{if} \quad g(x) = 1 \\ \mathbb{P}(Y = 0|X = x), & \textit{if} \quad g(x) = 0 \end{cases}$$

That is,

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \eta(x), & \textit{if} \quad g(x) = 1 \\ 1 - \eta(x), & \textit{if} \quad g(x) = 0 \end{cases}$$

Now, for fixed x, to have the maximal Probability $\mathbb{P}(Y = g(x)|X = x)$, we need to choose

Proof: Bayes Predictor in the BC Problem

Now, for a fixed x, consider

$$\mathbb{P}(Y = g(x)|X = x).$$

We have

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \mathbb{P}(Y = 1|X = x), & \textit{if} \quad g(x) = 1 \\ \mathbb{P}(Y = 0|X = x), & \textit{if} \quad g(x) = 0 \end{cases}$$

That is,

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \eta(x), & \textit{if} \quad g(x) = 1 \\ 1 - \eta(x), & \textit{if} \quad g(x) = 0 \end{cases}$$

Now, for fixed x, to have the maximal Probability $\mathbb{P}(Y = g(x)|X = x)$, we need to choose $\eta(x)$, if $\eta(x) \geq \frac{1}{2}$, or $1 - \eta(x)$, if $\eta(x) < \frac{1}{2}$ (in the case when $\eta = 0.5$, you can choose either one).

Proof: Bayes Predictor in the BC Problem

Now, for a fixed x, consider

$$\mathbb{P}(Y = g(x)|X = x).$$

We have

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \mathbb{P}(Y = 1|X = x), & \textit{if} \quad g(x) = 1 \\ \mathbb{P}(Y = 0|X = x), & \textit{if} \quad g(x) = 0 \end{cases}$$

That is,

$$\mathbb{P}(Y = g(x)|X = x) = \begin{cases} \eta(x), & \textit{if} \quad g(x) = 1 \\ 1 - \eta(x), & \textit{if} \quad g(x) = 0 \end{cases}$$

Now, for fixed x, to have the maximal Probability $\mathbb{P}(Y = g(x)|X = x)$, we need to choose $\eta(x)$, if $\eta(x) \geq \frac{1}{2}$, or $1 - \eta(x)$, if $\eta(x) < \frac{1}{2}$ (in the case when $\eta = 0.5$, you can choose either one). So, for maximal Probability, we need to have $g(x) = 1$ for $\eta(x) \geq \frac{1}{2}$, and $g(x) = 0$ otherwise. And this is exactly our g^*.

Notes

Let us summarize: the following Predictor

$$g^*(x) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1|X = x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

is a Bayes Predictor for our 0-1 Classification Problem.

Notes

Let us summarize: the following Predictor

$$g^*(x) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1|X = x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

is a Bayes Predictor for our 0-1 Classification Problem. Btw, the following is also a Bayes Classifier:

$$g^*(x) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1|X = x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Notes

Let us summarize: the following Predictor

$$g^*(x) = \begin{cases} 1, & \text{if } \mathbb{P}(Y=1|X=x) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

is a Bayes Predictor for our 0-1 Classification Problem. Btw, the following is also a Bayes Classifier:

$$g^*(x) = \begin{cases} 1, & \text{if } \mathbb{P}(Y=1|X=x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

So in this case Bayes Classifier is not unique.

Note

Note: We can express this in other form.

Note

Note: We can express this in other form. We know, from the Bayes Theorem, that

$$\eta(x) = \mathbb{P}(Y = 1|X = x) =$$

Note

Note: We can express this in other form. We know, from the Bayes Theorem, that

$$\eta(x) = \mathbb{P}(Y = 1|X = x) =$$

$$= \frac{\mathbb{P}(X = x|Y = 1) \cdot \mathbb{P}(Y = 1)}{\mathbb{P}(X = x|Y = 0) \cdot \mathbb{P}(Y = 0) + \mathbb{P}(X = x|Y = 1) \cdot \mathbb{P}(Y = 1)}$$

Note

Note: We can express this in other form. We know, from the Bayes Theorem, that

$$\eta(x) = \mathbb{P}(Y = 1 | X = x) =$$

$$= \frac{\mathbb{P}(X = x | Y = 1) \cdot \mathbb{P}(Y = 1)}{\mathbb{P}(X = x | Y = 0) \cdot \mathbb{P}(Y = 0) + \mathbb{P}(X = x | Y = 1) \cdot \mathbb{P}(Y = 1)}$$

Then the condition $\eta(x) \geq \frac{1}{2}$ can be written in the form

$$\mathbb{P}(X = x | Y = 1) \cdot \mathbb{P}(Y = 1) \geq \mathbb{P}(X = x | Y = 0) \cdot \mathbb{P}(Y = 0).$$

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$.

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$. Again we take the 0-1 Loss, i.e., the loss is 0, if we predict correctly, and 1 otherwise.

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$. Again we take the 0-1 Loss, i.e., the loss is 0, if we predict correctly, and 1 otherwise.

In the analogy, we denote

$$\eta_k(x) = \mathbb{P}(Y = k | X = x).$$

η_k is the **Posterior Class Probability** (for the class no. k).

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$. Again we take the 0-1 Loss, i.e., the loss is 0, if we predict correctly, and 1 otherwise.

In the analogy, we denote

$$\eta_k(x) = \mathbb{P}(Y = k | X = x).$$

η_k is the **Posterior Class Probability** (for the class no. k).

Then the following will be a Bayes Predictor (Classifier) for the K-class Classification Problem:

$$g^*(x) =$$

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$. Again we take the 0-1 Loss, i.e., the loss is 0, if we predict correctly, and 1 otherwise.

In the analogy, we denote

$$\eta_k(x) = \mathbb{P}(Y = k | X = x).$$

η_k is the **Posterior Class Probability** (for the class no. k).

Then the following will be a Bayes Predictor (Classifier) for the K-class Classification Problem:

$$g^*(x) = \underset{k=1,...,K}{argmax}\, \eta_k(x)$$

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$. Again we take the 0-1 Loss, i.e., the loss is 0, if we predict correctly, and 1 otherwise.

In the analogy, we denote

$$\eta_k(x) = \mathbb{P}(Y = k | X = x).$$

η_k is the **Posterior Class Probability** (for the class no. k).

Then the following will be a Bayes Predictor (Classifier) for the K-class Classification Problem:

$$g^*(x) = \underset{k=1,...,K}{argmax}\, \eta_k(x) = \underset{k=1,...,K}{argmax}\, \mathbb{P}(Y = k | X = x).$$

Generalization for K Classes

Assume we are solving now K-class Classification Problem, i.e., $\mathcal{Y} = \{1, 2, ..., K\}$. Again we take the 0-1 Loss, i.e., the loss is 0, if we predict correctly, and 1 otherwise.

In the analogy, we denote

$$\eta_k(x) = \mathbb{P}(Y = k | X = x).$$

η_k is the **Posterior Class Probability** (for the class no. k).

Then the following will be a Bayes Predictor (Classifier) for the K-class Classification Problem:

$$g^*(x) = \underset{k=1,...,K}{argmax}\, \eta_k(x) = \underset{k=1,...,K}{argmax}\, \mathbb{P}(Y = k | X = x).$$

And again, we can write this as

$$g^*(x) = \underset{k=1,...,K}{argmax}\; \mathbb{P}(X = x | Y = k) \cdot \mathbb{P}(Y = k).$$

Question, K-class Classification Problem

What will happen, if we will take the following Loss Function:

$$\ell(k, m) = |k - m| \ ?$$