Lecture 6 - 07-04-2020

(X, Y) We random variables drawn iid from D on $X \cdot Y \longrightarrow$ where D is fixed but unknown

Independence does not hold. We do not collect datapoints to an independent process.
Example: identify new article and i want to put categories. The feed is highly depend on what is happening in the world and there are some news highly correlated. Why do we make an assumption that follows reality? Is very convenient in mathematical term. If you assume Independence you can make a lot of process in mathematical term in making the algorithm.
If you have enough data they look independent enough. Statistical learning is not the only way of analyse algorithms —> we will see in linear ML algorithm and at the end you can use both statistical model s

1.1 Bayes Optimal Predictor

$$f^* : X \rightarrow Y$$
$$f^*(x) = argmin \, \mathbb{E} \left[\, \ell(y, \hat{y}) | X = x \, \right] \qquad \hat{y} \in Y$$

In general Y given X has distribution $D_y | X = x$
Clearly $\forall \ h \quad X \rightarrow Y$

$$\mathbb{E} \left[\, \ell(y, f^*(x)) | X = x \, \right] \leq \mathbb{E} \left[\, \ell(y, h(x) | X = x \, \right]$$
$$X, Y \qquad \mathbb{E} \left[\, Y | X = x \, \right] = F(x) \qquad \longrightarrow Conditional Expectation$$
$$\mathbb{E} \left[\, \mathbb{E} \left[\, Y | X \, \right] \right] = \mathbb{E}(Y)$$

Now take Expectation for distribution

$$\mathbb{E} \left[\, \ell(y, f^*(x)) \, \right] \leq \left[\, \mathbb{E}(\ell(y, h(x)) \, \right]$$

where risk is smaller in f^*
I can look at the quantity before
l_d Bayes risk $\longrightarrow$ Smallest possible risk given a learning problm

$$l_d(f^*) > 0 \qquad \textit{because } y \textit{ are still stochastic given } X$$

Learning problem can be complem $\rightarrow$ large risk

1.1.1 Square Loss

$$\ell(y, \hat{y} = (y - \hat{y})^2$$

I want to compute bayes optimal predictor
$\hat{y}, y \in \mathbb{R}$

$$f^*(x) = argmin\, \mathbb{E}\left[\,(y - \hat{y})^2 | X = x\,\right] = \qquad \hat{y} \in \mathbb{R}$$

we use $\qquad \mathbb{E}\,[\,X + Y\,] = \mathbb{E}[X]{+}\mathbb{E}[Y] = argmin\, \mathbb{E}\left[\,y^2 + \hat{y}^2 - 2 \cdot y \cdot \hat{y}^2 | X = x\,\right] =$

Dropping y^2 i remove something that is not important for $\hat{y}$

$$= argmin(\mathbb{E}\left[\,y^2 | X = x\,\right] + \hat{y}^2 - 2 \cdot \hat{y} \cdot \mathbb{E}\,[\,y | X = x\,]) =$$
$$= argmin(\hat{y}^2 - 2 \cdot \hat{y} \cdot \mathbb{E}\,[\,y | X = x\,]) =$$

Expectation is a number, so it's a constant
Assume $\boxdot = y^2$

$$argmin\,\left[\boxdot + \hat{y}^2 + 2 \cdot \hat{y} \cdot \mathbb{E}\,[\,Y | X = x\,]\right]$$

where red$G(\hat{y})$ is equal to the part between [...]

$$\frac{dG(\hat{y})}{d\hat{y}} = 2 \cdot \hat{y} - 2 \cdot \mathbb{E}\,[\,y | X = x\,] = 0 \quad \longrightarrow \quad \textit{So setting derivative to 0}$$

— DISEGNO OPT CURVE —

$G'(\hat{y}) = \hat{y}^2 - 2 \cdot b \cdot \hat{y}$

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

Square loss is nice because expected prediction is ...
In order to predict the best possibile we have to estimate the value given

data point.

$$\mathbb{E}\left[\, (y - f^*(x))^2 | X = x \,\right] =$$
$$= \mathbb{E}\left[\, (y - \mathbb{E}\left[\, y | X = x \,\right])^2 | X = x \,\right] = Var\left[\, Y | X = x \,\right]$$

1.1.2 Zero-one loss for binary classification

$Y = \{-1, 1\}$

$$\ell(y, \hat{y}) = I\{\hat{y} \neq y\} \qquad I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

If $\hat{y} \neq y$ true, indicator function will give us 1, otherwise it will give 0

$$D \quad on \qquad X \cdot Y \qquad D_x^* \quad D_{y|x} = D$$

$$D_x \qquad \eta : X \longrightarrow [\, 0, 1 \,] \qquad \eta = \mathbb{P}\,(y = 1 | X = x)$$

$$D \rightsquigarrow (D_x, \eta) \quad \longrightarrow \quad \textit{Distribution 0-1 loss}$$

$X \backsim D_x \quad \longrightarrow \quad$ *Where $\backsim$ mean "draw from" and D_x is marginal distribution*

$$Y = 1 \qquad \textit{with probability } \eta(x)$$

$$D_{y|x} = \{\eta(x), 1 - \eta(x)\}$$

Suppose we have a learning domain
— DISEGNO –
where η is a function of x, so i can plot it
η will te me $Prob(x) =$
η tells me a lot how hard is learning problem in the domain
$\eta(x)$ is not necessary continous
— DISEGNO —

$\eta(x) \in \{0, 1\} \qquad y$ is always determined by x

How to get f^* from the graph?

$$f^+ : X \to \{-1, 1\}$$
$$Y = \{-1, +1\}$$

— DISEGNO —

==================================

MANCA ROBAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

==================================

$$f^*(x) = argmin\, \mathbb{E}\,[\,\ell(y, \hat{y})|X = x\,] = \qquad \longrightarrow \hat{y} \in \{-1, +1\}$$

$$= argmin\, \mathbb{E}\,[\,I\{\hat{y} = 1\} \cdot I\{Y = -1\} + I\{\hat{y} = -1\} \cdot I\{y = 1\}\,|\,X = x\,] =$$

we are splitting wrong cases

$$= argmin\,(\,I\{\hat{y} = 1\} \cdot \mathbb{E}\,[\,I\{Y = -1\}|\,X = x\,] + I\{\hat{y} = -1\} \cdot \mathbb{E}\,[\,I\{y = 1\}\,|\,X = x\,]\,) = \quad ⋇$$

We know that:

$$\mathbb{E}\,[\,I\{y = -1\}\,|\,X = x\,] = 1 \cdot \mathbb{P}(\hat{y} = -1|X = x) + 0 \cdot \mathbb{P}(y = 1|X = x) =$$
$$\mathbb{P}(x = -1|X = x) = 1 - \eta(x)$$

$$⋇ = argmin\,(\,I\{\hat{y} = 1\} \cdot (1 - \eta(x)) + I\{\hat{y} = -1\} \cdot (\eta(x)\,)\,)$$

where Blue colored $I\{...\} = 1°$ and Orange $I\{...\} = 2°$

I have to choose -1 or +1 so we will **remove one of the two (1° or 2°)**

It depend on $\eta(x)$:

- If $\eta(x) < \frac{1}{2} \quad \longrightarrow \quad$ kill 1°
- Else $\eta(x) \geq \frac{1}{2} \quad \longrightarrow \quad$ kill 2°

$$f^*(x) = \begin{cases} +1 & if\, \eta(x) \geq \frac{1}{2} \\ -1 & if\, \eta(x) < \frac{1}{2} \end{cases}$$

1.2 Bayes Risk

$$\mathbb{E}\,[\,I\{y \neq f^*(x)\}\,|\,X = x\,] = \mathbb{P}(y \neq f^*(x)|X = x)$$

$$\eta(x) \geq \frac{1}{2} \quad \Rightarrow \quad \hat{y} = 1 \quad \Rightarrow \quad \mathbb{P}(y \neq 1|X = x) = 1 - \eta(x)$$

$$\eta(x) < \frac{1}{2} \quad \Rightarrow \quad \hat{y} = -1 \quad \Rightarrow \quad \mathbb{P}(y \neq 1|X = x) = \eta(x)$$

Conditiona risk for 0-1 loss is:

$$\begin{aligned}\mathbb{E}\,[\,\ell(y, f^*(x))\,|\,X = x\,] \quad &= \quad I\{\eta(x) \geq \frac{1}{2}\} \cdot (1 - \eta(x)) + I\{\eta(x) < \frac{1}{2}\} \cdot \eta(x) = \\ &= min\,\{\eta(x), 1 - \eta(x)\}\end{aligned}$$

$$\mathbb{E}\,[\,\ell, f^*(x)\,] = \mathbb{E}\,[\,min\,\{\eta(x), 1 - \eta(x)\}\,]$$

`bayesrisk.jpg`

Conditional risk will be high aroun the half so min between the two is around the half since the labels are random i will get an error near 50%.
My condition risk will be 0 in the region in the bottom since label are going to be deterministic.