Motivation

Supervised Learning: Learning from a randomly sampled batch of labeled data

What is the Objective of the Learner?

Dataset (tuple of tuples) $\xrightarrow{\text{Learner (function)}}$ Predictor = Model = Hypothesis (function)

$$D = \left((\vec{X}_1, Y_1), (\vec{X}_2, Y_2), \ldots, (\vec{X}_n, Y_n)\right) \in (\mathcal{X} \times \mathcal{Y})^n$$

$$(\vec{X}_i, Y_i) \sim P_{\vec{X}, Y}$$

independent for all $i \in \{1, \ldots, n\}$

D n feature-label pairs

$\mathcal{X}$ set of d-dimensional features

$\mathcal{Y}$ set of labels/targets

$$f : \mathcal{X} \to \mathcal{Y}$$

f a function from features to labels

Ex $f(x) = 200x + 100$, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$$

$\mathcal{A}$ a function from datasets to predictors

Ex: $\mathcal{A}(D) = \hat{f}$ where $\hat{f} : \mathcal{X} \to \mathcal{Y}$

$$\hat{f}(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i \in \{1, \ldots, n\} \\ & \text{Pick } i \text{ to be lowest index} \\ 0 & \text{otherwise} \end{cases}$$

Setting:

We are given a random dataset of size n

$$D = ((\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$$

where $(\vec{X}_i, Y_i) \sim \mathbb{P}_{\vec{X},Y}$ are independent for all $i \in \{1, \dots, n\}$

$\vec{X}_i$ feature vector

Y_i label or target

We will always assume the features are vectors.

Ex (of features and labels/targets):

$\vec{X}_i \in \mathbb{R}^3$ # of rooms, # of floors, age of a house

$Y_i \in \mathbb{R}$ price

$\vec{X}_i \in \mathbb{R}^2$ amount of chemical 1, amount of chemical 2 in a wine

$Y_i \in \{0, 1\}$ type of wine

$\vec{X}_i \in \mathbb{R}^{400}$ pixel value of a $20 \times 20 = 400$ pixel image

$Y_i \in \{cat, dog, bird\}$ type of animal

What is a feature and what is a label is a design choice. Usually a feature is info that is easy to gather. And the label is hard, which is why you want to predict it

Objective (Informal)

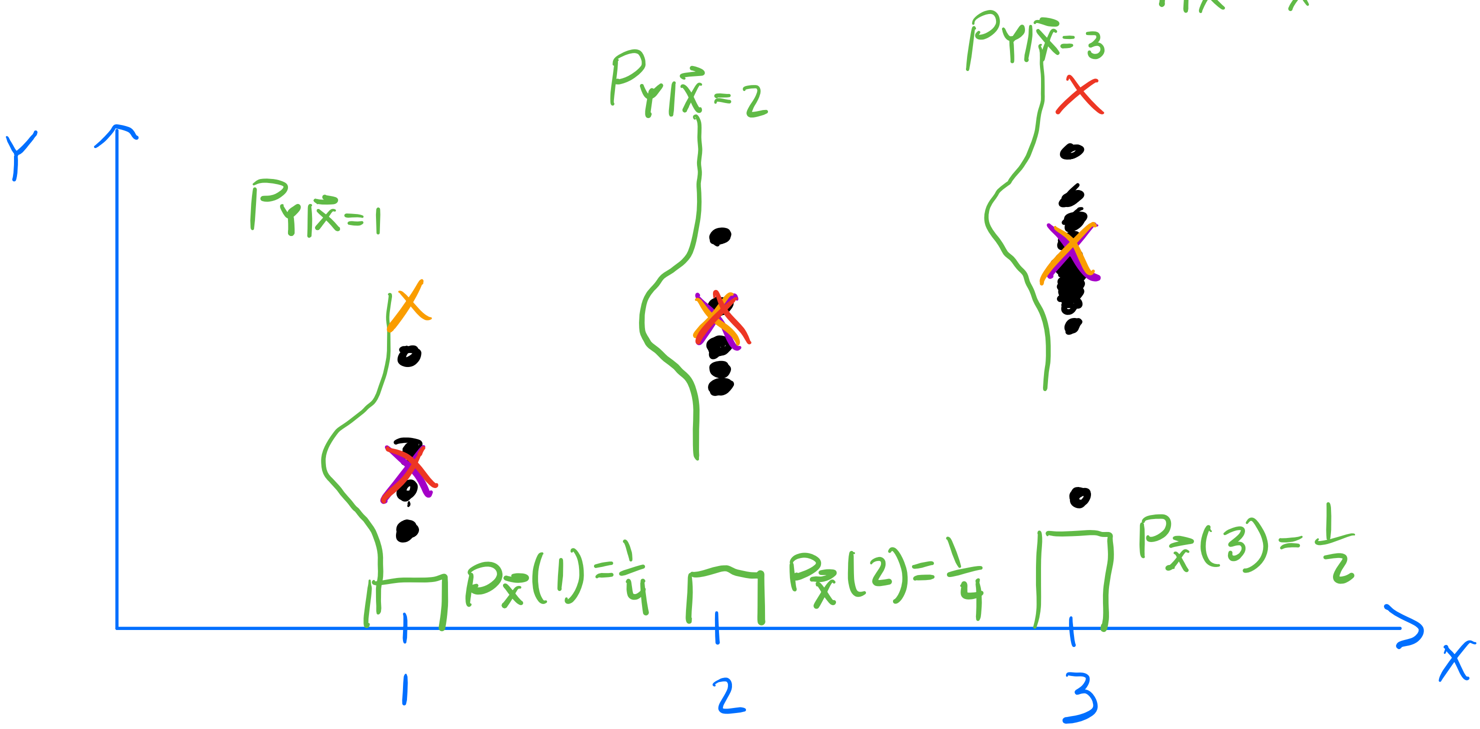
Define a learner $\mathcal{A}(D) = \hat{f}$

this is unknown

Such that for any new $(X, Y) \sim \mathbb{P}_{\vec{X},Y}$

$= \mathbb{P}_{Y|\vec{X}} \mathbb{P}_{\vec{X}}$

$\hat{f}(X) = Y$



$\hat{f} : \{1, 2, 3\} \Rightarrow \mathbb{R}$

$\ell : \mathcal{Y} \times \mathcal{Y} \Rightarrow [0, \infty)$

$\ell(\hat{f}(x), y) = |\hat{f}(x) - y|$

loss, cost, error

$\mathbb{E}[\,|\hat{f}(X) - Y| \mid X = 1]$

$\mathbb{E}[\,|\hat{f}(X) - Y| \mid X = 2]$

$\mathbb{E}[\,|\hat{f}(X) - Y| \mid X = 3]$

Scalarize

$$\sum_x p_X(x) = 1$$

$$\sum_{x \in \{1,2,3\}} \mathbb{E}\left[|\hat{f}(\vec{X}) - Y| \,\middle|\, X = x \right] p_X(x)$$

$$= \mathbb{E}_{\mathbb{P}_{\vec{X}}}\left[\mathbb{E}_{\mathbb{P}_{Y|\vec{X}}}\left[\underbrace{|\hat{f}(\vec{X}) - Y|}_{Z} \,\middle|\, X \right] \right] \qquad \mathbb{E}[\mathbb{E}[Z|X]] = \mathbb{E}[Z]$$

$$= \mathbb{E}\left[|\hat{f}(\vec{X}) - Y| \right]$$

$$= \mathbb{E}\left[\underbrace{\ell(\hat{f}(\vec{X}), Y)}_{\in [0,\infty)} \right]$$

Expected loss

Risk

$$= \sum_{x \in \{1,2,3\}} \left(\int_{\mathbb{R}} |\hat{f}(x) - y| \, p_{Y|\vec{X}=x}(y|x) \, dy \right) p_X(x)$$

$$=: L(\hat{f})$$

Regression: $Y \in \mathcal{Y}$ represent something with order

(usually $\mathcal{Y}$ is $\mathbb{R}$ or an interval)

Ex: house prices, stock prices, weather prediction

We use:

$$\ell(\hat{y}, y) = |\hat{y} - y|$$ absolute value loss

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

Classification:

$Y \in \mathcal{Y}$ represent something without order

Ex: type of wine, type of image, type of disease

We use:

$$\ell(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{if } \hat{y} \neq y \end{cases}$$ 0-1 loss

Objective (more formal)

Define a learner $\mathcal{A}(D) = \hat{f}$

this is unknown

Such that for any new $(X, Y) \sim \mathbb{P}_{\vec{X}, Y}$

$= \mathbb{P}_{Y|\vec{X}} \mathbb{P}_{\vec{X}}$

$L(\mathcal{A}(D))$ is small

D is random!

If D changes then $\hat{f}$

$L(\mathcal{A}(D_1)), L(\mathcal{A}(D_2)), \ldots$

$D_i \in (\mathcal{X} \times \mathcal{Y})^n$

Scalarize

$$\int_{(\mathcal{X} \times \mathcal{Y})^n} L(\mathcal{A}(D)) \underbrace{w(D)}_{= P_D} \, dD$$

$\mathcal{A}(D) = \hat{f}$

$$= \mathbb{E}[L(\mathcal{A}(D))] = \mathbb{E}\big[\mathbb{E}[\ell(\mathcal{A}(D)(\vec{X}), Y) \mid D]\big]$$

Objective (more formal)

Define a learner $\mathcal{A}(D) = \hat{f}$

Such that for any new $(X, Y) \sim \mathbb{P}_{\vec{X},Y}$

this is unknow

$\mathbb{E}[L(\mathcal{A}(D))]$ is small

$= \mathbb{P}_{Y|\vec{X}} \mathbb{P}_{\vec{X}}$

pick $\hat{f}$ that minimizes $L(\hat{f})$

$$L(\hat{f}) = \mathbb{E}[\ell(\hat{f}(X), Y)]$$

Need to know $P_{\vec{X},Y}$

$\mathcal{A}$ = Empirical Risk Minimizer (ERM)

input: D

Function Class

choice reflects prior knowledge

Estimation: Use D to estimate $L(f)$

for all $f \in \mathcal{F} \subseteq \{f \mid f: \mathcal{X} \to \mathcal{Y}\}$

call the estimate $\hat{L}(f)$

Optimization: pick $\hat{f} \in \mathcal{F}$ that minimizes $\hat{\tilde{L}}(\hat{f})$

No Free Lunch Theorem (Informal)

If $\mathcal{A} = ERM$ uses $\mathcal{F} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$

and $\mathbb{E}[L(\mathcal{A}(D))] \leq \frac{1}{4}$

Then $n \geq \frac{|\mathcal{X}|}{2}$

Ex: function class $\mathcal{F} = \{\text{all lines}\}$