Regression: What is the best $\hat{f}$?

Statistical Learning: (X, Y) have some joint distribution $p(x, y)$

and $Y = f^*(X)$

unknown

* Want $\hat{f}$ that is an estmate of f^*
  so that $Y \approx \hat{f}(X)$

* We construct $\hat{f}$ by defing a Loss $L: \mathbb{R}^p \to \mathbb{R}$

$$\hat{f} = \underset{f}{\operatorname{argmin}} \overbrace{\mathbb{E}[L(f)]}^{\text{Risk}}$$

So far $L(f) = (Y - f(X))^2$

Reality: training $\{(x_n, y_n)\}$

minimize empircal risk

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y_n - f(x_n))^2}_{\text{MSE}}$$

What is theoretically $\hat{f}$? (for sq. Loss)

Quantity of interest:

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

$$= \iint (y - f(x))^2 \, p(x, y) \, dx \, dy$$

density fn

Iterated Exp.

$$\mathbb{E}[X] = \mathbb{E}_X[\mathbb{E}_{Y|X}[X]]$$

$$\mathbb{E}[XY] = \mathbb{E}_X \mathbb{E}_{Y|X}[XY]$$

density fn

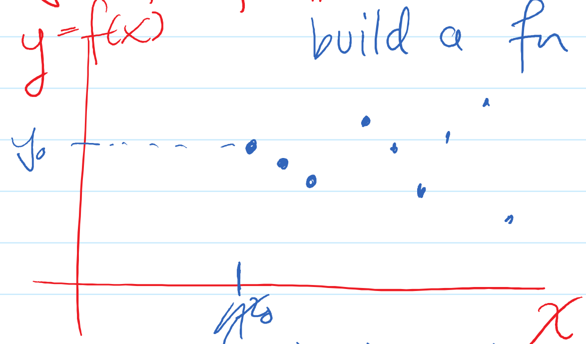
$$= \iint (y-f(x))^2\, p(y|x)\, p(x)\, dx\, dy$$

$$(*) = \int \left[\int (y-f(x))^2\, p(y|x)\, dy \right] p(x)\, dx = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[(Y-f(X))^2 \right] \right]$$

$$\boxed{\hat{f} = \arg\min_f \mathbb{E}[L(f)]}$$

picking a fn $f: \mathbb{R}^p \to \mathbb{R}$

build a fn



building $\hat{f}$ can be done one x at a time

$\{f(x)\}_x$

Aside

just a large table

x	$f(x)$
1	3
1.1	.4
1.01	7.3
π	e

$$(*) \int \underbrace{\left[\int (y-f(x))^2\, p(y|x)\, dy \right]}_{A(x)} p(x)\, dx = \boxed{\int A(x)\, p(x)\, dX}$$

$$A(x) = \mathbb{E}\left[(Y-f(X))^2 \mid X = x \right]$$

depends on $f(x) \in \mathbb{R}$

want to minimize (choose f to minimize this)

(1) I can choose $f(x)$ independently for each x

(chos... minimize ...)

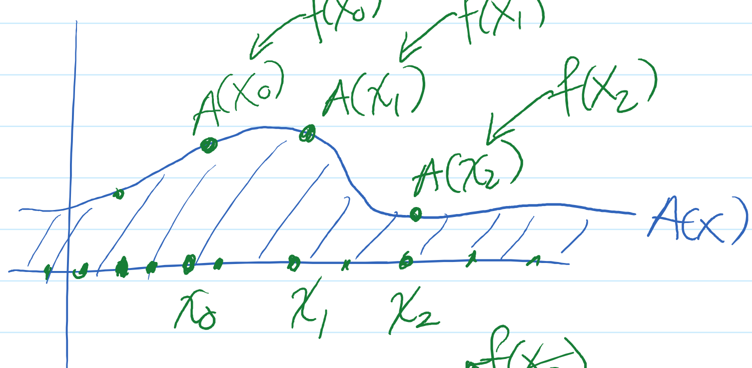
② So $A(x)$ are independently set (according to chosen $f(x)$) at each x

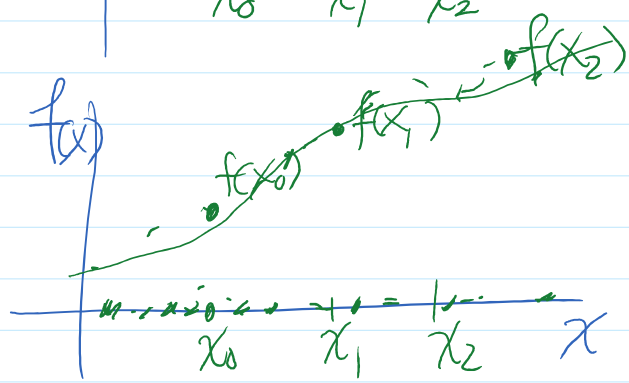
Want to minimize $\boxed{\int A(x)p(x)dx}$

$p(x)$ given can't choose

choice of $f(x)$

Way to minimize is to minimize $A(x)$ independently at each x by choosing $\boxed{f(x)}$ to minimize $A(x)$





(*) $$\hat{f}(x) = \arg\min_{f(x)} A(x) = \arg\min_{f(x)} \underbrace{E\left[(Y-f(x))^2 \mid X=x\right]}_{A(x)}$$

Fact:

$$\hat{a} = \arg\min_a \mathbb{E}\left[(Z-a)^2\right]$$

plug in $\mathbb{E}[Z]$

$$\Rightarrow \mathbb{E}\left[(Z-a)^2\right] = \mathbb{E}\left[Z^2 - 2aZ + a^2\right]$$

$$= \mathbb{E}[Z^2] - 2a\,\mathbb{E}[Z] + a^2$$

$$\frac{\partial}{\partial a}\left[\cdots\right] = -2\,\mathbb{E}[Z] + 2a = 0$$

$$\boxed{\hat{a} = \mathbb{E}[Z]}$$

We said

$$\hat{f}(x) = \arg\min_{f(x)} \mathbb{E}\left[(\underbrace{Y}_{Z} - \underbrace{f(x)}_{a})^2 \mid X = x\right]$$

$$\boxed{\hat{f}(x) = \mathbb{E}\left[Y \mid X = x\right]} \quad (*)$$

Fun fact: $L(f) = |Y - f(X)|$

$$\hat{f}(x) = \text{Median}(Y \mid X = x)$$

How does this relate?

$$\boxed{\hat{f}(x) = \mathbb{E}[Y|X=x]}$$

In practice: we approximate.

<u>How?</u> ① Collect some data $\{(x_n, y_n)\}$

② Approx. $\mathbb{E}[Y|X=x]$

<u>How?</u>

<u>Way 1:</u> $\hat{f}(x)$ = avg value of y_ns for x_ns near x
(lets use K nearest neighbors to x)

$$= \frac{1}{K} \sum_{n: x_n \in N_K(x)} y_n$$

(KNN)

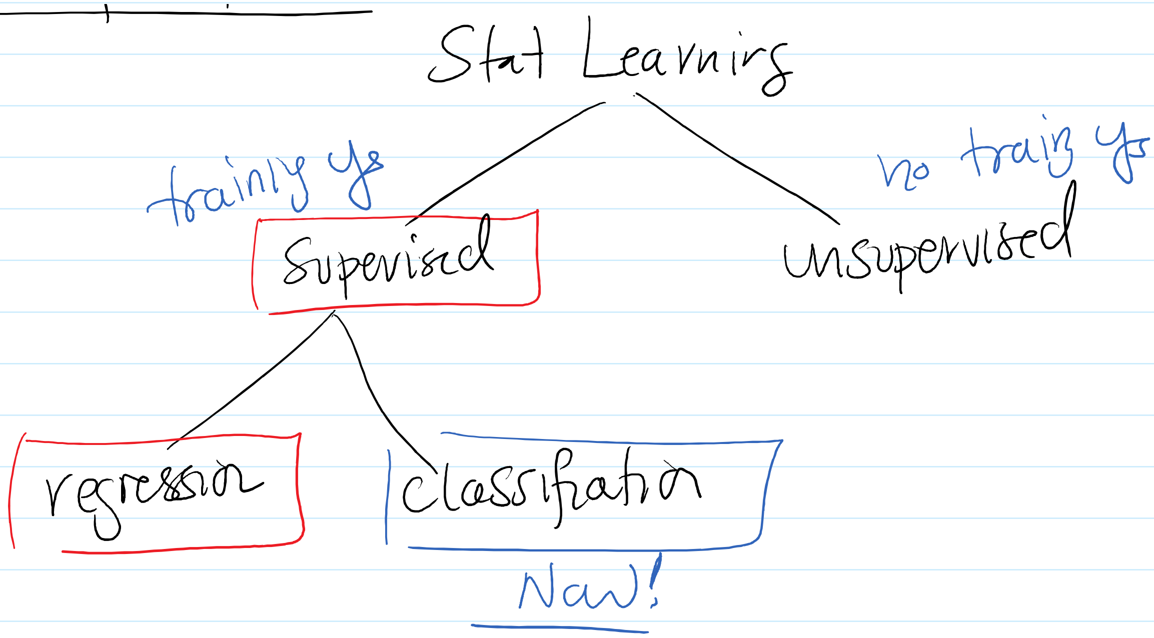
<u>Way 2:</u> Assume some linear structure of $\mathbb{E}[Y|X=x] = x^T\beta$

I estimate βs as $\hat{\beta}$

$\hat{f}(x) = \hat{\beta}^T x$ (Linear regression)

<u>Classification:</u>

Stat Learning



Classification: (Categorical ys)

- predict song title from sounds — multi-class
- predict tumor benign/malignant — 2-class
- predict protein class: (1) α-helix (2) β-helix (3) random coil — 3-class

---

Setup: $\{(y_n, x_n)\}$: $x_n \in \mathbb{R}^p$; $y_n \in \mathcal{C}$

$$\mathcal{C} = \{c_1, c_2, c_3, \ldots, c_K\}$$

$\mathcal{C}$ set of K classes

Goal: find $\hat{f}$ so that $\hat{y} = \hat{f}(x) \approx y$

Optimal $\hat{f}$?

Need a loss for classification: "0-1 loss"

$$L(f) = \mathbb{1}(f(x) \neq Y) = \begin{cases} 0 & f(x) = Y \\ 1 & f(x) \neq Y \end{cases}$$

$$\hat{f}(x) = \underset{f(x)}{\arg\min}\ \mathbb{E}[\overbrace{\mathbb{1}(Y \neq f(x)}^{\text{loss}} \mid X = x]$$

Aside: $\mathbb{E}[\mathbb{1}(X \in A)] = \int \mathbb{1}(x \in A)\, p(x)\, dx$

$$= \int_A p(x)\, dx = \mathbb{P}(X \in A)$$

$$= \underset{f(x)}{\arg\min}\ P(Y \neq f(x) \mid X = x)$$

$$= \underset{f(x)}{\arg\min}\ 1 - \underbrace{P(Y = f(x) \mid X = x)}_{\text{maximize this}}$$

$$\boxed{\hat{f}(x) = \underset{c \in \mathcal{C}}{\arg\max}\ \mathbb{P}(Y = c \mid X = x)}$$

Bayes classifier

Idea: see! x

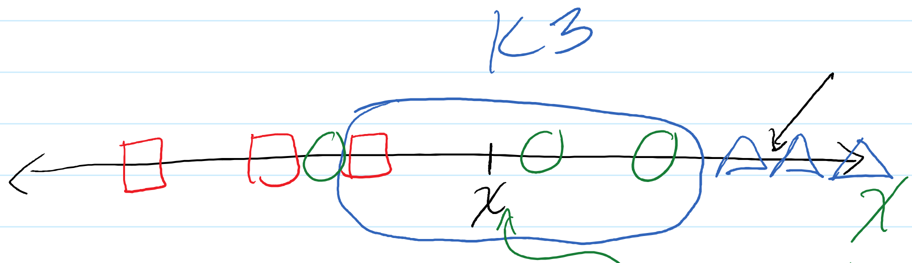
class probs

$\mathbb{P}(Y=c_1 \mid X=x),\ \mathbb{P}(Y=c_2 \mid X=x),\ \mathbb{P}(Y=c_3 \mid X=x), \ldots$

pick largest.

KNN classification

classes: □ ○ △



$\hat{f}(x) =$ majority class among K nearest neighbors

$\hat{f}(x) = \bigcirc$

$$p_c = \mathbb{P}(Y=c \mid X=x) \approx \frac{1}{K} \overbrace{\sum_{n:\, x_n \in N_K(x)} \mathbb{1}(y_n = c)}^{\#\text{ class } c}$$

= % of K nearest neighbors of class c

pick $\hat{f}(x)$ = class w/ largest.