Regression: Sq. err loss $L(f) = (Y - f(X))^2$

Classification: **Setup:** $\{(x_n, y_n)\}$

$x_n \in \mathbb{R}^p$ and $y_n \in \mathcal{C}$

Goal: find $\hat{f}$ so that $Y \approx \hat{f}(X)$

$$\underbrace{\mathcal{C} = \{c_1, c_2, \ldots, c_K\}}_{\text{discrete set of classes.}}$$

Loss for classification: 0-1 loss

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

Empirical /Risk Minimization

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

Shown:

$$\hat{f}(x) = \arg\min_{f(x)} \mathbb{E}[\overbrace{(Y - f(X))^2}^{L(f)} \mid X = x]$$

(for Sq. Loss)

$$= \arg\min_{f(x)} \mathbb{E}[L(f) \mid X = x]$$

For classification $Y \in \mathcal{C}$ so $f(x) \in \mathcal{C}$

$$\hat{y} = \hat{f}(x) = \underset{f(x)}{\operatorname{argmin}}\ E\left[\mathbb{1}(Y \neq f(x)) \mid X = x\right]$$

$$= \underset{c \in \mathcal{C}}{\operatorname{argmin}}\ E\left[\mathbb{1}(Y \neq c) \mid X = x\right]$$

Aside: $E[\mathbb{1}(X \in A)] = \mathbb{P}(X \in A)$

$$\int \mathbb{1}(x \in A) f(x)\, dx = \int_A f(x)\, dx$$

$$= \underset{c \in \mathcal{C}}{\operatorname{argmin}}\ \mathbb{P}(Y \neq c \mid X = x)$$

$$= \underset{c}{\operatorname{argmin}}\ 1 - \underbrace{\mathbb{P}(Y = c \mid X = x)}_{\max}$$

$$\hat{y} = \hat{f}(x) = \underset{c}{\operatorname{argmax}}\ \mathbb{P}(Y = c \mid X = x)$$

Bayes' classifier

what we do:

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

and then we choose class that maximizes among all these.

Bayes' Rate: $\max_{c} \mathbb{P}(Y = c \mid X = x)$

prob. I am correct

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

from Bayes'

'' classifier ✓

Bayes' Classifier: says look at $\mathbb{P}(Y=c \mid X=x)$

① **discriminative models** — model $Y|X$

→ estimate $P(Y=c|x)$ directly

↳ KNN classification, logistic regression, classification trees

② **generative models**

→ Bayes' rule:

$$\mathbb{P}(Y=c \mid x) = \frac{P(X=x \mid Y=c)\,\mathbb{P}(Y=c)}{\mathbb{P}(X=x)}$$

doesn't depend on class

$$\propto \underbrace{P(X=x \mid Y=c)}_{\text{class conditional prob.}}\mathbb{P}(Y=c)$$

So model ① $X|Y$ and ② Y

↳ LDA/QDA, naive bayes

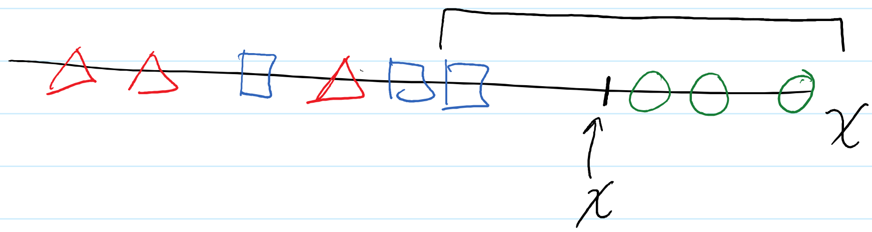
KNN Classification: (discriminative method) model $Y|X$

classes: △ □ ○

'' = 4

$K = 4$

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



$\hat{p}_{\triangle} = 0$

$\hat{p}_{\square} = 1/4$

$\hat{p}_{\bigcirc} = 3/4 \leftarrow$ choose

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

% of KNNs that are class c

$$\mathbb{P}(Y=c \mid X=x) \approx \hat{p}_c = \frac{1}{K} \sum_{n:\, x_n \in N_K(x)} \mathbb{1}(y_n = c)$$

$$\mathbb{1}(y_n = c) = \begin{cases} 1 & y_n = c \\ 0 & y_n \neq c \end{cases}$$

and then $\hat{y} = \arg\max_c \hat{p}_c$

Evaluation: train/validation/test

Error rate: % of misclassified examples

$\{(x_n, y_n)\}$ and on $\hat{f}$ then

$$\text{error} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(y_n \neq \hat{y}_n)$$

$\hat{y}_n = \hat{f}(x_n)$

Accuracy: $1 -$ error

$=$ % of correctly classified

(*) Can look at these per class (*)

Confusion Matrix :

Ex.

3 classes

Fill in w/ the number in each bin

truth \ predicted	1	2	3
1	10	5	7
2	3	2	13
3	2	7	1

→ accuracy = sum of diag / sum of mtx $= \frac{13}{50}$

→ error = sum of dig / sum of mtx $= \frac{37}{50}$

→ accuracy of class 1 = $10/22$