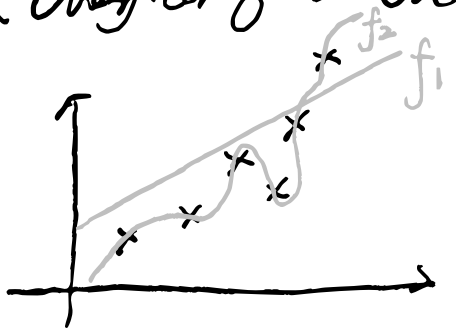
§ Introduction

Statistical Learning : $Y = f(X) + \varepsilon$

※ Overfitting & Underfitting



f_1 or f_2 ?

※ Training data & Testing data

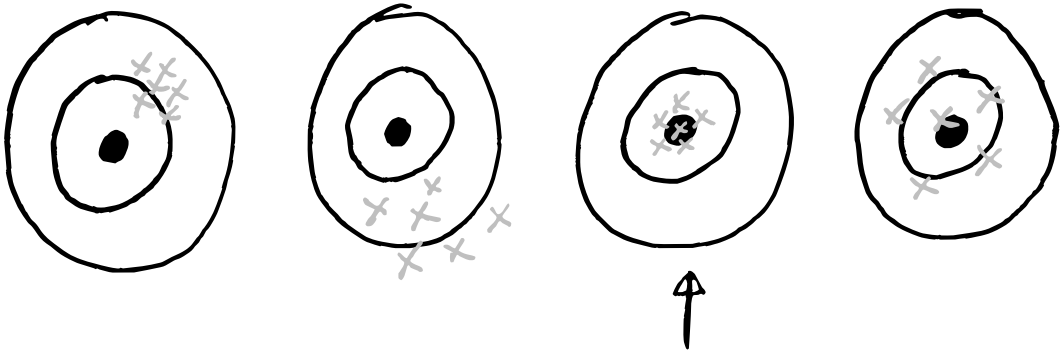
$\begin{cases} \text{traditionally} : 80\% \text{ vs. } 20\% \text{ etc.} \\ \text{DL era} : 99\% \text{ vs. } 1\% \text{ etc.} \end{cases}$

※ Evaluation of fit

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \begin{cases} \text{training} \\ \text{testing} \end{cases}$$

$$\mathbb{E}\left[\left(y_0 - f(x_0)\right)^2\right] = Var(\hat{f}(x_0)) + \left[bias(\hat{f}(x_0))\right]^2 + Var(\varepsilon)$$

※ Bias-Variance tradeoff



※ Interpretation vs. Prediction

※ Regression vs. Classification
(Y is continuous) (Y is discrete)

※ Supervised vs. Unsupervised
(X, Y) (only X → clustering)

§ Logistic Regression (LR)

$$y = \begin{cases} 1, & \text{if event occurs} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{P}(y=1) = p \rightarrow \mathbb{E}(y) = p$$

※ How to model "p"?

$$f(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \in (-\infty, +\infty)$$

so, $f(p) = p$ is problematic

$$\rightarrow \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$\rightarrow p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

※ Prediction: $\hat{y_i} = \begin{cases} 1, & \text{if } p_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$

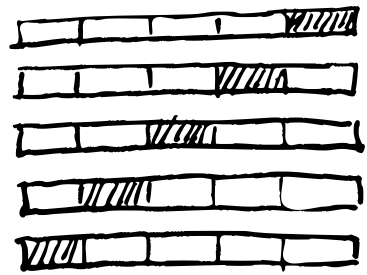
※ Estimation: MLE + Newton-Raphson

§ Cross validation, Evaluation measures

※ Cross validation (CV)

- waste less training data
- test on 100% of the data

e.g. 5-fold CV



- training
- testing

e.g. leave-one-out (LOO)

(n-1) training & 1 testing

※ Evaluation measures

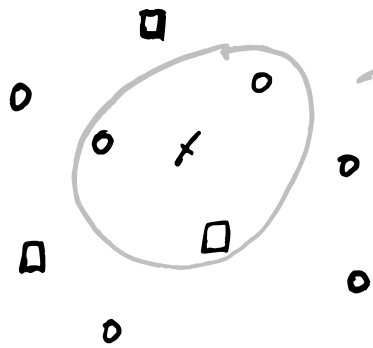
confusion matrix & $F = \frac{2 \times TP}{2 \times TP + FP + FN}$

	Predicted True	Predicted False
Actual True	TP	FN
Actual False	FP	TN

§ k-Nearest Neighbor

※ $\mathbb{P}(Y=j \mid X=x_0) = \frac{1}{k} \sum_{i \in N_0} \mathbb{1}\{y_i = j\}$

where $\begin{cases} x_0 : \text{a new observation} \\ N_0 : \text{the set of the nearest } K \text{ observations} \end{cases}$



$\mathbb{P}(x=\circ) = \frac{2}{3}$

$\mathbb{P}(x=\square) = \frac{1}{3}$

< when $k=3$ >

※ How to break the classification tie?

1) at random
2) increase k until tie is broken
3) use 1NN as tie breaker

※ Distances : Euclidean & Cosine

§ Naïve Bayes

※ Bayes' rule: $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A) \times \mathbb{P}(A)}{\mathbb{P}(B)}$

※ $\mathbb{P}(y|x_1, \cdots, x_n) = \dfrac{\mathbb{P}(x_1|y)\mathbb{P}(x_2|y)\cdots\mathbb{P}(x_n|y)\mathbb{P}(y)}{\mathbb{P}(x_1)\mathbb{P}(x_2)\cdots\mathbb{P}(x_n)}$

$\rightarrow \mathbb{P}(y|x_1, \cdots, x_n) \propto \mathbb{P}(y) \prod_{i=1}^{n} \mathbb{P}(x_i|y)$

※ Laplace Smoothing

eliminate probilities of 0 or 1

※ Example: stolen or not?

		Y	N
1.° color	Red	3/5	$\boxed{2/5}$
	Yellow	2/5	3/5
2.° type	Sports	4/5	2/5
	SUV	1/5	$\boxed{3/5}$
3.° origin	Domestic	2/5	$\boxed{3/5}$
	Imported	3/5	2/5

{ Red, SUV, Domestic

$\mathbb{P}(Y) = \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{2}$

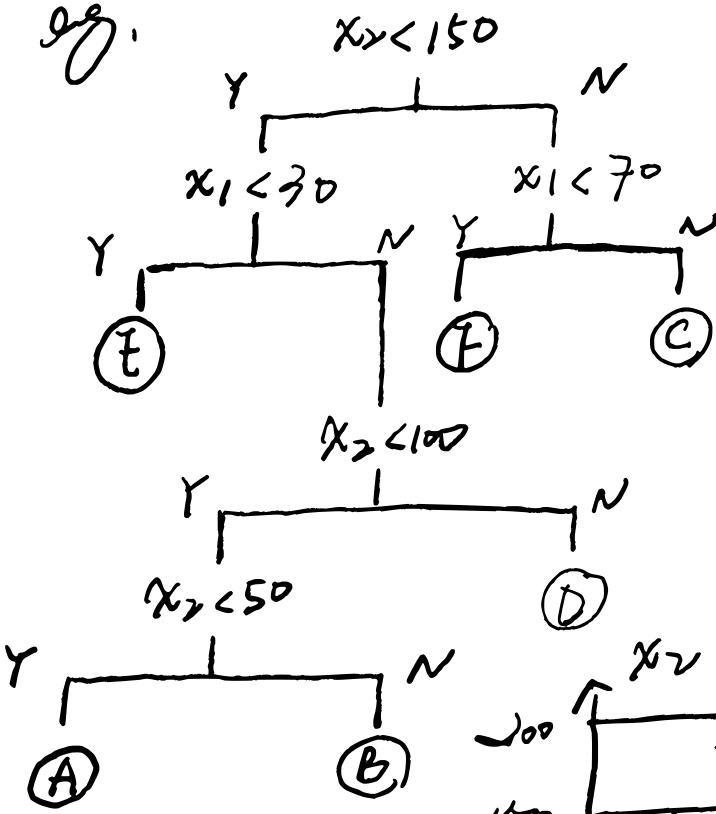
$\mathbb{P}(N) = \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{1}{2}$

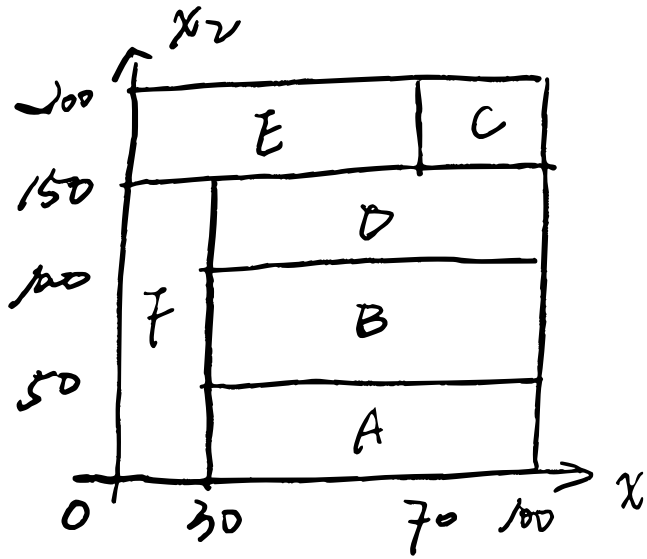
$\Rightarrow$ No!

§ Trees & Random forests

※ Classification And Regression Trees (CART)

e.g.





"Rectangles"

※ full tree vs. pruning

Criteria $\begin{cases} \text{classification error} \\ \text{Gini index} \\ \text{cross-entropy} \end{cases}$

※ Bootstrap : sample with replacement

※ Bagging (Bootstrap AGgregation)

→ $\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$

→ reduce variance

→ classifiers : $\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$

※ Random forests "Majority vote"

for b = 1 to B :

draw a bootstrap sample

grow a tree $\begin{cases} \text{select } m \text{ out of } p \text{ variables} \\ \text{choose the best split} \end{cases}$

output ensemble

※ Out-of-bag (OOB)

Each bootstrap sample on average only uses 2/3 of the observations

→ the remaining 1/3 observations can be used to estimate the test error

→ train / test split is not needed

§ Boosting

※ Forward stagewise additive modelling

$f_0(x) = 0$

for $m = 1$ to M :

$\min \sum_{i=1}^{N} \mathcal{L}[f_{m-1}(x) + \beta_m b(x; \gamma_m)]$

$f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$

※ Exponential Loss → Adaboost

Binomial likelihood → Logit-boost

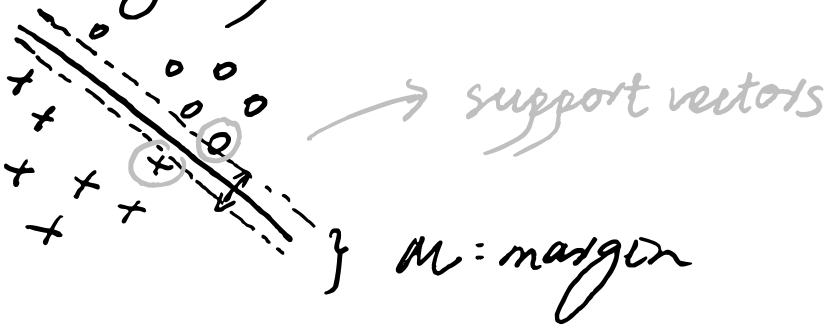
※ Regularization

→ fit less aggressively to avoid overfitting

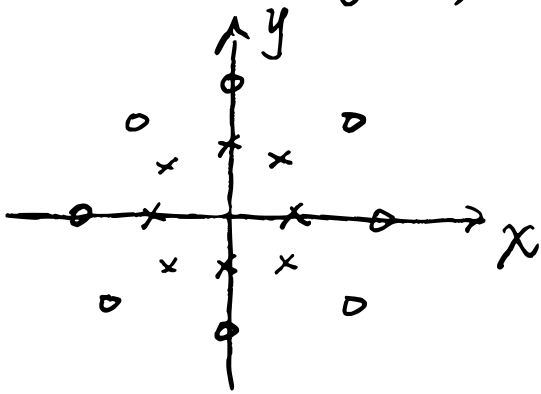
$\begin{cases} \text{add penalty to loss function} \\ \text{shrinking} \end{cases}$

§ Support Vector Machines (SVM)

※ Linearly Separable



※ Not Linearly Separable



$$\begin{cases} \times : x^2+y^2 = 1 \\ \circ : x^2+y^2 = 2 \end{cases}$$

$$\rightarrow z = \sqrt{x^2+y^2} \qquad \begin{cases} \times : z = 1 \\ \circ : z = \sqrt{2} \end{cases}$$

"Linearly separable"