BOOSTING

COMBINE WEAK LEARNERS THAT WERE TRAINED SEQUENTIALLY AND PREDICTS BASED ON THE WEIGHTS OF EACH MODEL (WEIGHT CALCULATED BASED ON MODEL PERFORMANCE).

GRADIENT BOOSTING

$dY(0) = Y - \text{MEAN}(Y)$

FOR $K = 1 : (\#\text{OF TREES} / \#\text{ OF ITERATIONS})$:

$\text{LEARNER}(K) = \text{TRAIN\_REGRESSOR}(X, dY(K-1))$

$dY(K) = dY(K-1) - \alpha(K) * \text{PREDICT}(\text{LEARNER}(K), X)$

"TRAIN OVER THE RESIDUAL"

BAGGING

* GENERATE N DIFFERENT BOOTSTRAP TRAINING SAMPLE WITH REPLACEMENT
* TRAIN ALGORITHM ON EACH BOOTSTRAPPED SAMPLE
* COMBINE THEM ALL USING MAJORITY VOTE / MEAN

RELATIVE VARIABLE IMPORTANCE

THE MEASURE IS MADE BASED ON # TIMES A VARIABLE IS SELECTED FOR SPLITTING AND WEIGHTED BY THE IMPROVEMENT TO THE MODEL AS A RESULT OF EACH SPLIT, THEM AVERAGED OVER ALL TREES.

CONFUSION MATRIX

POSITIVE CLASS = 0

PREDICT \ OBSERVED	P	N
P	TP	FP
N	FN	TN

$$\text{SENSITIVITY} = \frac{TP}{TP+FN} = \text{RECALL}$$

$$\text{SPECIFICITY} = \frac{TN}{TN+FP}$$

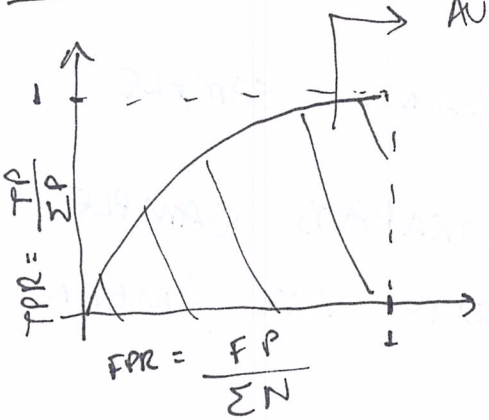
$$\text{PRECISION} = \frac{TP}{TP+FP}$$

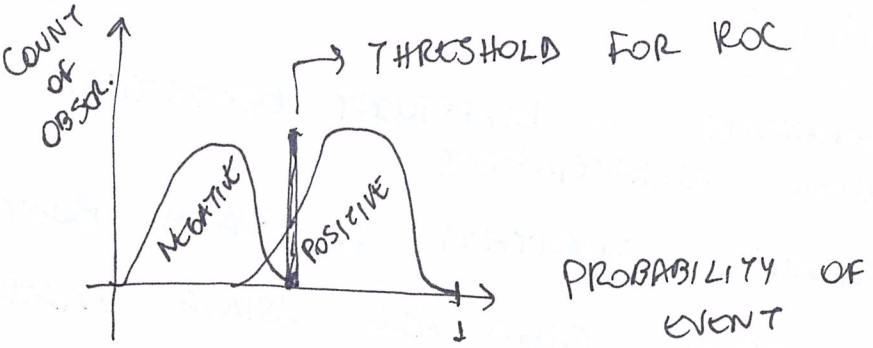
$$\text{F1 SCORE} = 2 \cdot \frac{\text{PRECISION} * \text{RECALL}}{\text{PRECISION} + \text{RECALL}} = \frac{2}{\frac{1}{\text{PRECISION}} + \frac{1}{\text{RECALL}}}$$

"HARMONIC MEAN"

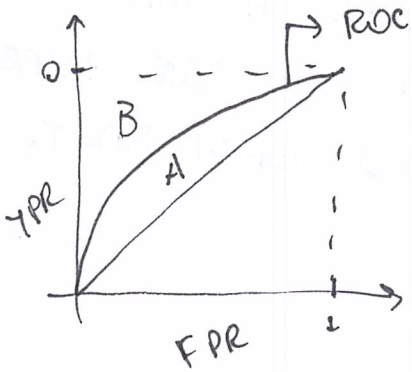
$$\text{ACCURACY} = \frac{TP+TN}{P+N}$$

ROC CURVE





GINI COEFFICIENT (INEQUALITY COEFFICIENT)

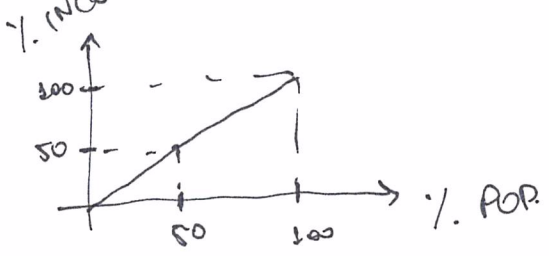


$$\text{GINI} = \frac{A}{A+B}$$

HIGHER A => BETTER MODEL

HIGHER INEQUALITY

REAL-WORLD EXAMPLE:



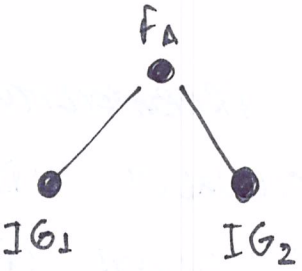
"PERFECT DISTRIBUTION OF MONEY OVER POPULATION"

CLASSIFICATION TREES

FOR ALL FEATURES TESTS INFORMATION GAIN USING ENTROPY:

$$IG = -\sum_{i=1}^{J} P_i \cdot \log_2 P_i$$

P_i = PROPORTION OF CLASS i ON NODE



$$IG_T = IG_1 + IG_2$$

(FOR FEATURE A THE INFORMATION GAIN WILL BE THE SUM OF $IG_1 + IG_2$)

REGRESSION TREES

AT EACH ITERATION FOR EACH FEATURE x_k FIND OPTIMAL S:

$$\min_S \left[MSE(y \mid x_k < s) + MSE(y \mid x_k \geq s) \right]$$

(*s IS THE [redacted] CUT OFF)

* FOR BOTH METHODS: VARIABLE IMPORTANCE GENERALLY BE COMPUTED BASED ON CORRESPONDING REDUCTION OF PREDICTIVE ACCURACY WHEN THE PREDICTOR OF INTEREST IS REMOVED OR SOME MEASURE OF DECREASE OF NODE IMPURITY.

LINEAR REGRESSION

LINEAR APPROACH TO MODELLING RELATIONSHIP BETWEEN SCALAR RESPONSE TO EXPLANATORY VARIABLES

"ORDINARY LEAST SQUARES"

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n \cdot X_n$$

LOGISTIC REGRESSION

LOG-ODDS OF PROBABILITY OF AN EVENT IS A LINEAR COMBINATION OF EXPLANATORY VARIABLES

"MAXIMUM LIKELIHOOD ESTIMATION"

$$P(Y=1) = \frac{1}{1+e^{-x}}$$

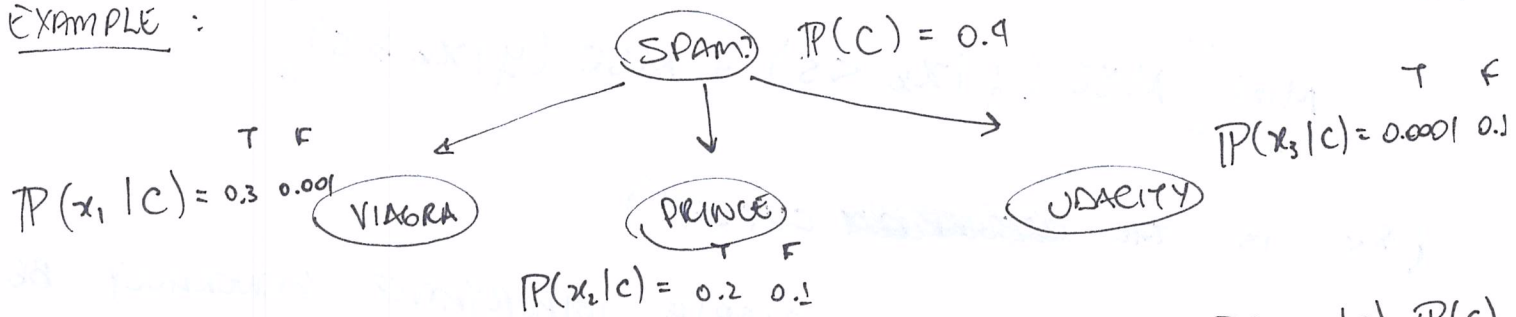
* VARIABLE IMPORTANCE FOR REGRESSION CAN BE SET BASED ON THE COEFFICIENTS ONLY IF THE FEATURES ARE NORMALIZED (CONTINUOUS) OR THEY ARE DISCRETE.

NAIVE BAYES CLASSIFICATION

BASED ON BAYES THEOREM WITH CONDITIONAL PROBABILITY OF EVENT TO PREDICT. ASSUMES x_i CONDITIONALLY INDEPENDENT OF EVERY OTHER FEATURE x_j $(i \neq j)$ GIVEN CATEGORY C_k =>

$$\mathbb{P}(C_k \mid x_1, \dots, x_n) \propto \mathbb{P}(C_k) \cdot \prod_{i=1}^{n} \mathbb{P}(x_i \mid C_k)$$

EXAMPLE:



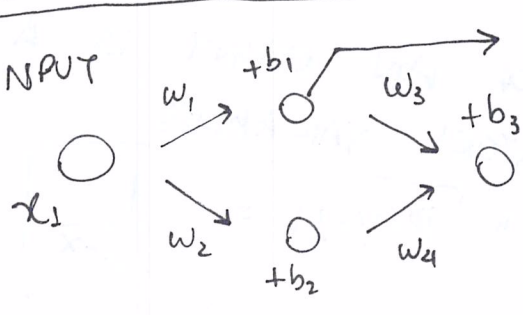
$$\mathbb{P}(C \mid x_1 = T, x_2 = F, x_3 = F) \propto \mathbb{P}(x_1 = T \mid C) \cdot \mathbb{P}(x_2 = F \mid C) \cdot \mathbb{P}(x_3 = F \mid C) \cdot \mathbb{P}(C)$$

RANDOM FOREST

SELECTING BAGGING SAMPLES FOR EACH TREE CHOOSE RANDOM FEATURES ($\sqrt{D}$).

* REDUCES VARIANCE BUT RANGE (AS ALL TREE MODELS) IS LIMITED.

NEURAL NETWORKS (ARTIFICIAL NEURAL NETWORKS)



$$\sigma(z = w_1 \cdot x_1 + b_1)$$

$\sigma(z) :=$ SIGMOID, TANH, RELU,

SIGMOID := $\sigma(z) = \frac{1}{1+e^{-z}}$; $\sigma(z) \in (0,1)$

TANH := $\sigma(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} = \tanh(z)$; $\sigma(z) \in (-1,1)$

RELU := $\sigma(z) = \max(0,z)$; $\sigma(z) \in [0, \infty)$.

SOFTMAX := $\sigma_i(\vec{z}) = \frac{e^{z_i}}{\sum_{j=1}^{J} e^{z_j}}$, $i = 1, \dots, J$; $\sigma(\vec{z}) \in (0,1)$

K-NN (K - NEAREST NEIGHBORS)

REGRESSION OR CLASSIFICATION OF K NEAREST NEIGHBORS BASED ON DISTANCE FUNCTIONS:

EUCLIDEAN : $\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$

MANHATTAN : $\sum_{i=1}^{k} |x_i - y_i|$

MINKOWSKI : $\left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q}$

K-MEANS

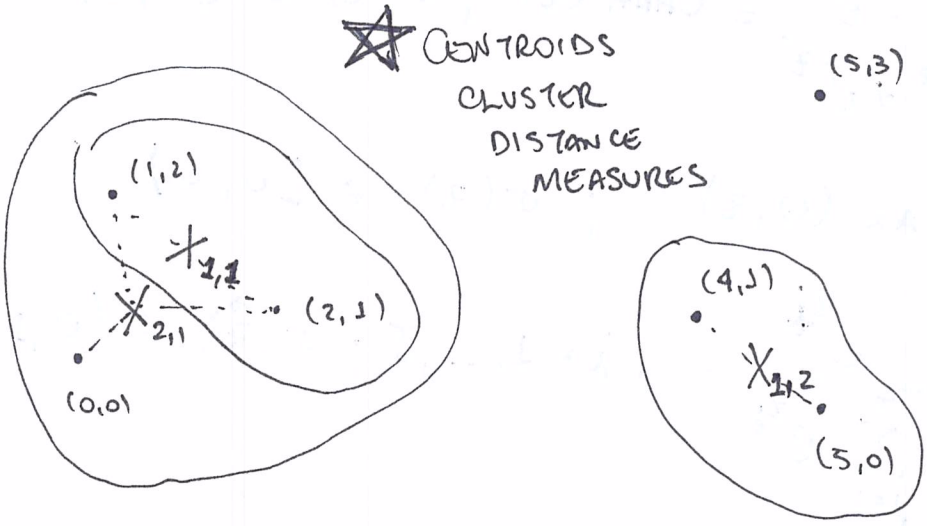
0. PLACE CENTROIDS AT RANDOM LOCATIONS (K CENTROIDS)
1. FIND NEAREST CENTROID TO EACH OBSERVATION
2. ASSIGN OBSERVATION TO CLOSER CLUSTER
3. CALCULATE NEW CENTROID

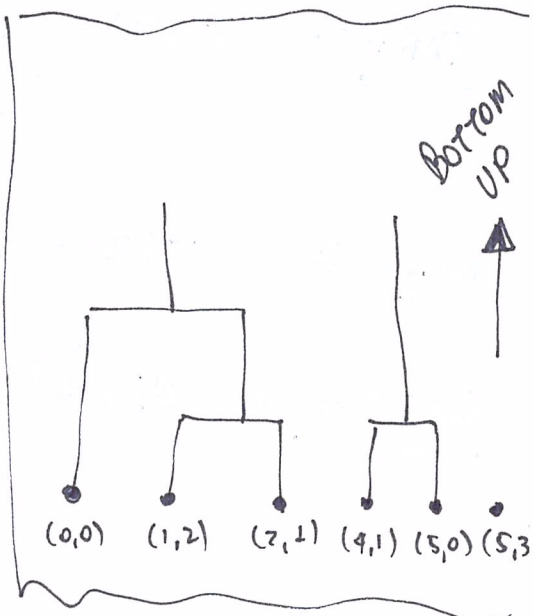
(k)

* REPEAT UNTILL DIFF. FROM PREVIOUS CENTROID IS MINIMUM (DIFFERENCE FROM DISTANCES)

→ K-MEANS ONLY USES EUCLIDEAN DISTANCE

HIERARCHICAL CLUSTERING





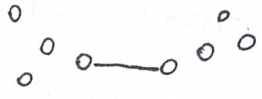
~~DISTANCES~~ DISTANCES TO BE USED:

EUCLIDEAN, MANHATTAN, MINKOWSKI, MAHALANOBIS := $\sqrt{(a-b)^T S^{-1} (a-b)}$

* S IS THE COVARIANCE MATRIX

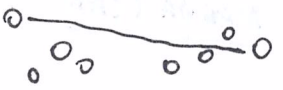
→ SINGLE LINKS: $D(C_1, C_2) = \min D(x_1, x_2)$

DISTANCE BETWEEN CLOSEST ELEMENTS IN CLUSTERS



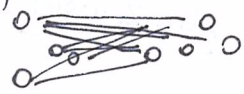
→ COMPLETE LINKS: $D(C_1, C_2) = \max D(x_1, x_2)$

DISTANCE BETWEEN FARTHEST ELEMENTS IN CLUSTERS



→ AVERAGE LINKS: $D(C_1, C_2) = \frac{1}{|C_1|} \frac{1}{|C_2|} \sum_{x_1} \sum_{x_2} D(x_1, x_2)$

AVERAGE OF ALL PAIRWISE DISTANCES

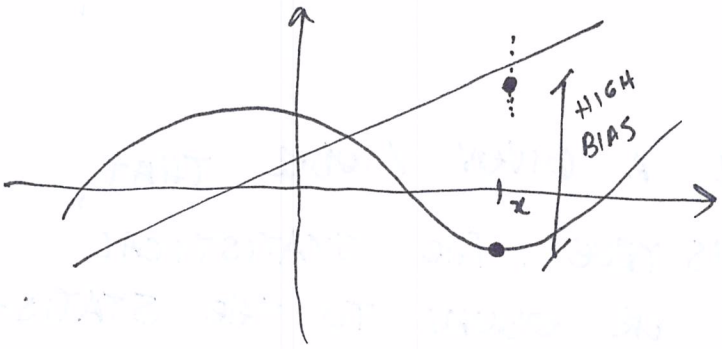


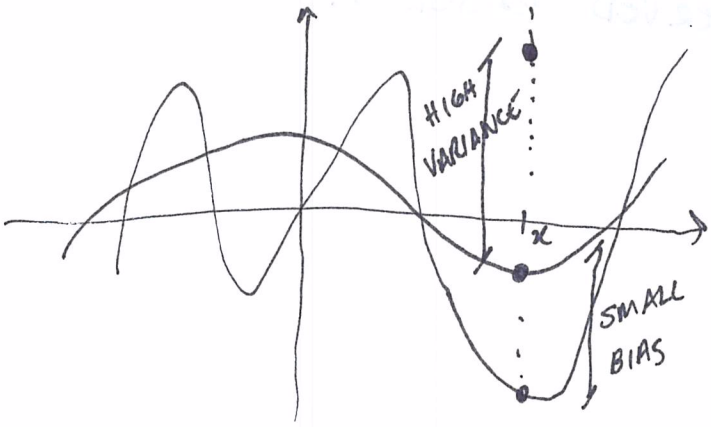
→ CENTROIDS: $D(C_1, C_2) = D\left[\left(\frac{1}{|C_1|} \sum_{x_1} \vec{x}\right), \left(\frac{1}{|C_2|} \sum_{x_2} \vec{x}\right)\right]$

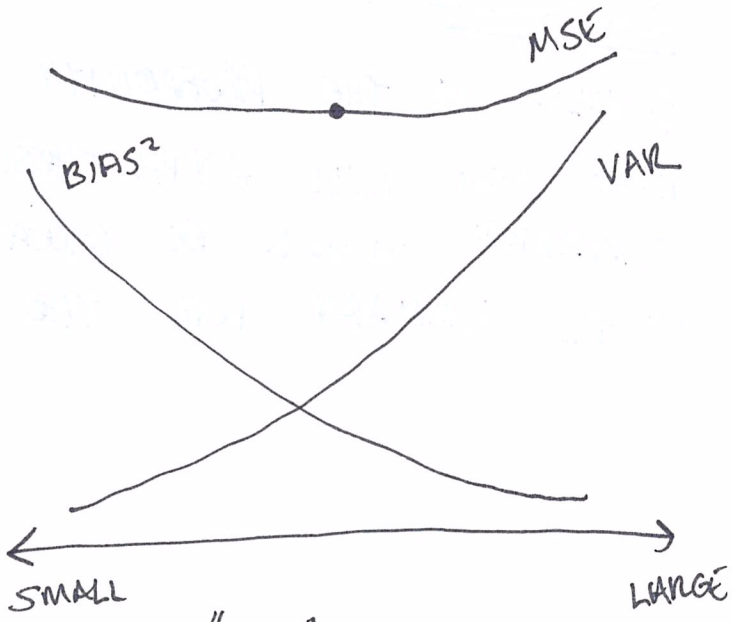
DISTANCE BETWEEN CENTROIDS (MEANS) OF TWO CLUSTERS

BIAS - VARIANCE DECOMPOSITION

- BIAS IS THE ERROR FROM ERRONEOUS ASSUMPTIONS IN THE MODEL
- VARIANCE IS THE ~~ERROR~~ ERROR FROM SENSITIVITY TO SMALL FLUTUATIONS IN THE TRAINING SET.







$$MSE(\hat{\theta}) = BIAS^2(\hat{\theta}) + VAR(\hat{\theta})$$

- BIAS ↑ = UNDERFITTING
- VAR ↑ = OVERFITTING

CROSS-VALIDATION

PREVENTS OVERFITTING BECAUSE MODEL THAT FITS RANDOM NOISE ON TRAINING DATA WON'T PERFORM GOOD ON VALIDATION DATASET.

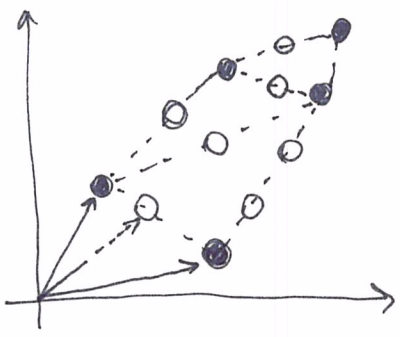
K-FOLD : SEPARATES DATA ON K FOLDS, TRAINING WILL BE K-1 FOLDS AND VALIDATION WILL BE 1

LOOCV : LEAVE-ONE-OUT CROSS-VALIDATION IS THE SAME AS K-FOLD BUT $K = N$.

SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

→ CREATES NEW "SYNTHETIC" OBSERVATIONS

- IDENTIFY FEATURE VECTOR AND NEAREST NEIGH.
- TAKE THE DIFF. BETWEEN TWO
- MULTIPLY DIFF. BY RANDOM BETWEEN 0 AND 1
- IDENTIFY NEW POINT ON LINE SEGMENT BY ADDING RANDOM NUMBER TO FEATURE VECTOR
- REPEAT @



P-VALUE

P-VALUE IS THE PROBABILITY FOR A GIVEN MODEL THAT, WHEN THE NULL HYPHOTHESIS IS TRUE, THE STATISTICAL SUMMARY WOULD BE GREATER OR EQUAL TO THE STATISTICAL SUMMARY FOR THE OBSERVED RESULTS.