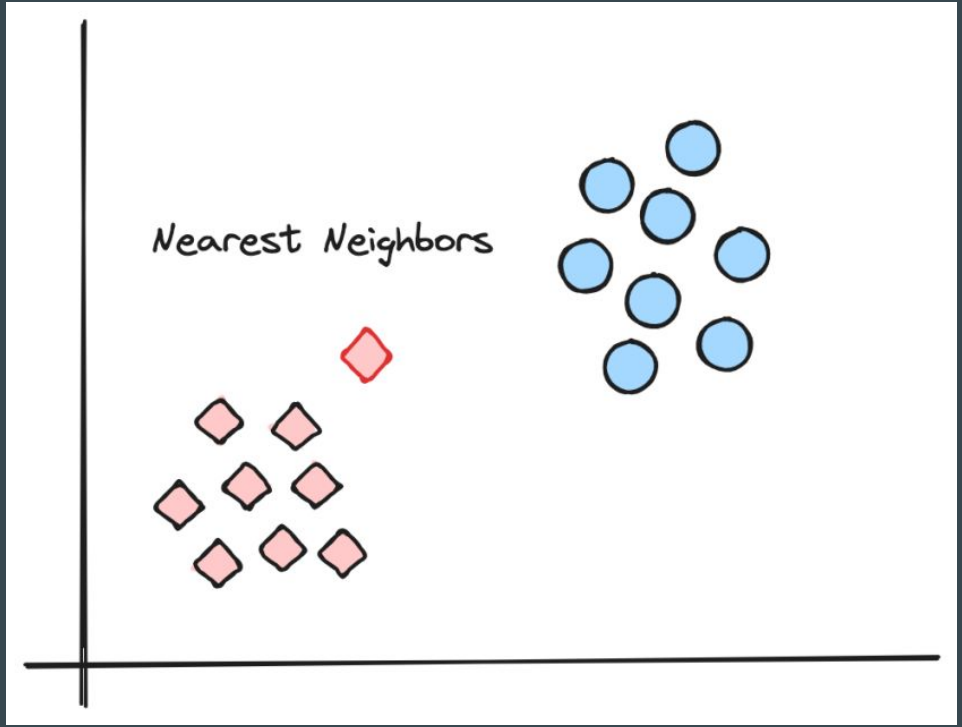
Locally Optimized Feature Weighted Ball Tree kNN - LOFWBT-kNN

•••

Dylan Tallis and Anna Wisniewski

Intro - kNN Algorithm

- Calculate distance from instance to all points
- Pick k closest and simple majority vote for predicted class



Related Work

James Keller: FKNN

- Distance weighing

Kumbure and Luukka: FWM-LMFKNN

- Feature weighting (MI)
- Minkowski distance
- Local means

Vahedifar et al.: IMKNN

- MI
- Shapley values

Pan et al.: DC-LAKNN

- Locally adaptive k

Rajani et al.:

- Ball Trees

Blood Donor Dataset

- 748 instances, 4 features, predicting future blood donation
- UCI ML Dataset
- Used as a benchmark by several other papers
- Numeric data

Variable Name	Role	Type	Description
Recency	Feature	Integer	months since last donation
Frequency	Feature	Integer	total number of donations
Monetary	Feature	Integer	total blood donated in c.c.
Time	Feature	Integer	months since first donation
Donated_Blood	Target	Binary	whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood)

Mammogram Dataset

- 961 instances, 6 features, class benign/malignant
- UCI ML dataset
- Mostly categorical (integer encoded)
- Missing values

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
BI-RADS	Feature	Integer				yes
Age	Feature	Integer	Age			yes
Shape	Feature	Integer				yes
Margin	Feature	Integer				yes
Density	Feature	Integer				yes
Severity	Target	Binary				no

Shapley Values

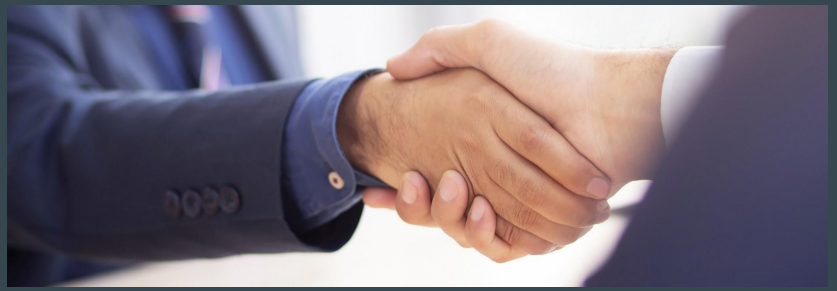
- Naive Bayes trained on dataset, Shapley values used to find feature contribution to classification
- Evaluates how much presence of feature improves accuracy compared to absent

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

Mutual Information

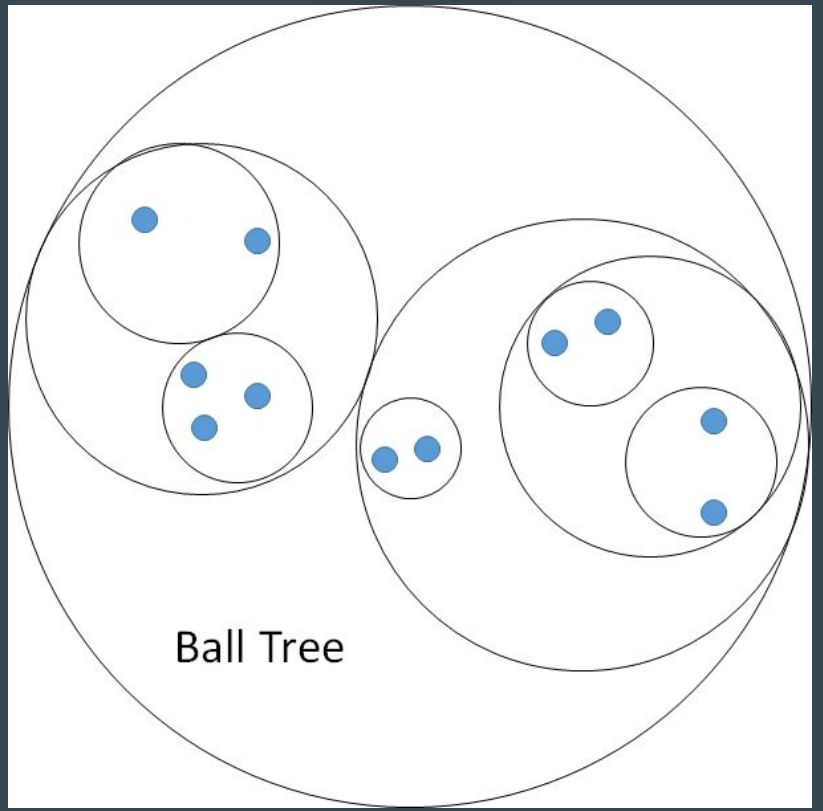
- Compares probability distribution of feature to joint of feature and class
- Compares probability of occuring together to expected if independent



$$\text{Mutual Information} = \sum_{y \in Y} \sum_{x \in X} p(x,y) \cdot log \frac{p(x,y)}{p(x)p(y)}$$

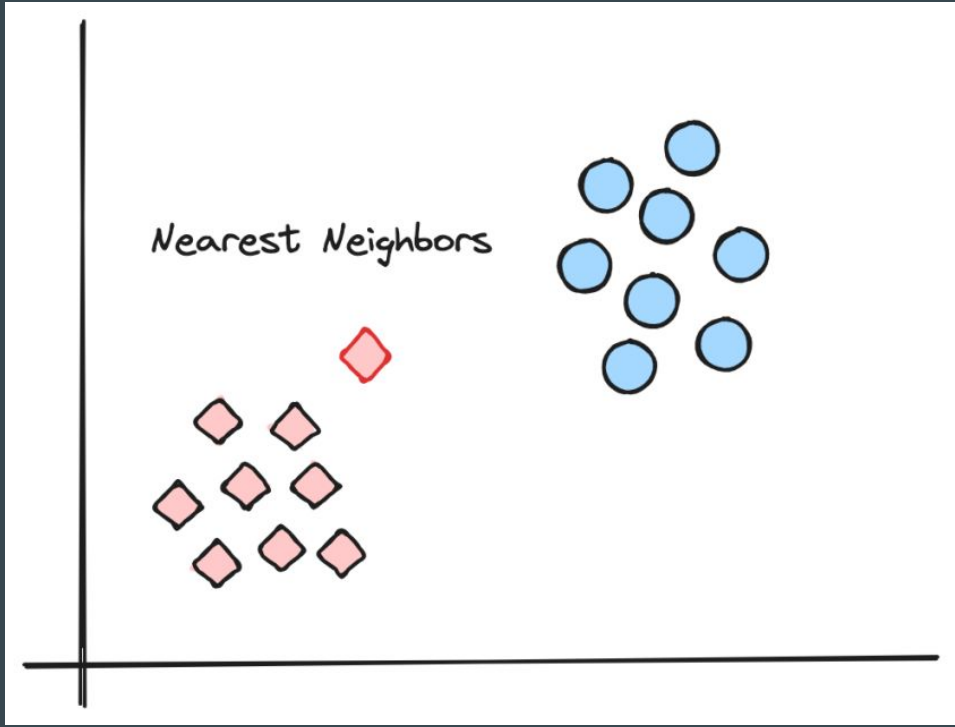
Ball Tree

- Partitions dataset to make searches more efficient
- O(logN) vs O(N)
- Splits recursively starting at pivot point
- Separates points within median distance from pivot and those outside



Local k Selection

- Based on consistency of classes in neighborhood
- Picks smallest k with high consistency



Results - Accuracies

	LOFWB-KNN	KNN	FWM-LMFKNN	FKNN
Mammogram	80.83	79.79	77.31	77.2
Blood	78.67	77.65	74.67	73.33

Results

Mammogram

Model	Precision	Recall
LOFWB-KNN	0.833	0.773
KNN	0.830	0.753
FKNN	0.791	0.742

Blood

Model	Precision	Recall
LOFWB-KNN	0.467	0.226
KNN	0.348	0.258
FKNN	0.320	0.258

Discussions

- Localized Naive Bayes
- Higher dimensional data
- Different weighting strategies
- Unbalanced data

Thanks! Any Questions?

...