به نام خدا

کارگاه علم داده با پایتون پیشرفته

جلسه چهارم: Classification یا طبقه بندی K نزدیک ترین همسایه (KNN)

مدرس :

مهرناز جلیلی

دانشجو کارشناسی ارشد علم داده ها

دانشگاه شهید بهشتی

Classification

Intro

- Understand Classification
- Understand different methods such as KNN, Decision Trees, Logistic Regression and SVM
- Apply on datasets
- Evaluate

Classification

Intro

- Supervised
- Categorizing unknown items in classes
- Target is categorical with discrete values (called classifier)
- Binary: 2 values vs Multi Class

Classification

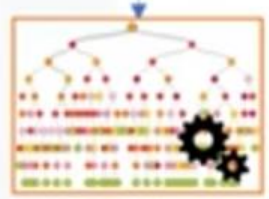
Intro

Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

Age	Sex	BP	Cholesterol	Na	K	Drug
36	F	LOW	HIGH	0.697	0.069	

Categorical Variable

Modeling



Classifier

Classification

Intro

- Loan (age, income, load size, previous records, ...)
- Churn (age, address, income, equip, data usage, calls, ...)
- Spam / Important email
- Handwriting/Speech recognition
- Biometric identification

Classification

Intro

- Decision Trees (ID3, C4.5, C5.0)
- Naive Bayes
- Linear Discriminant Analysis
- K-Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines

Classification

KNN

-

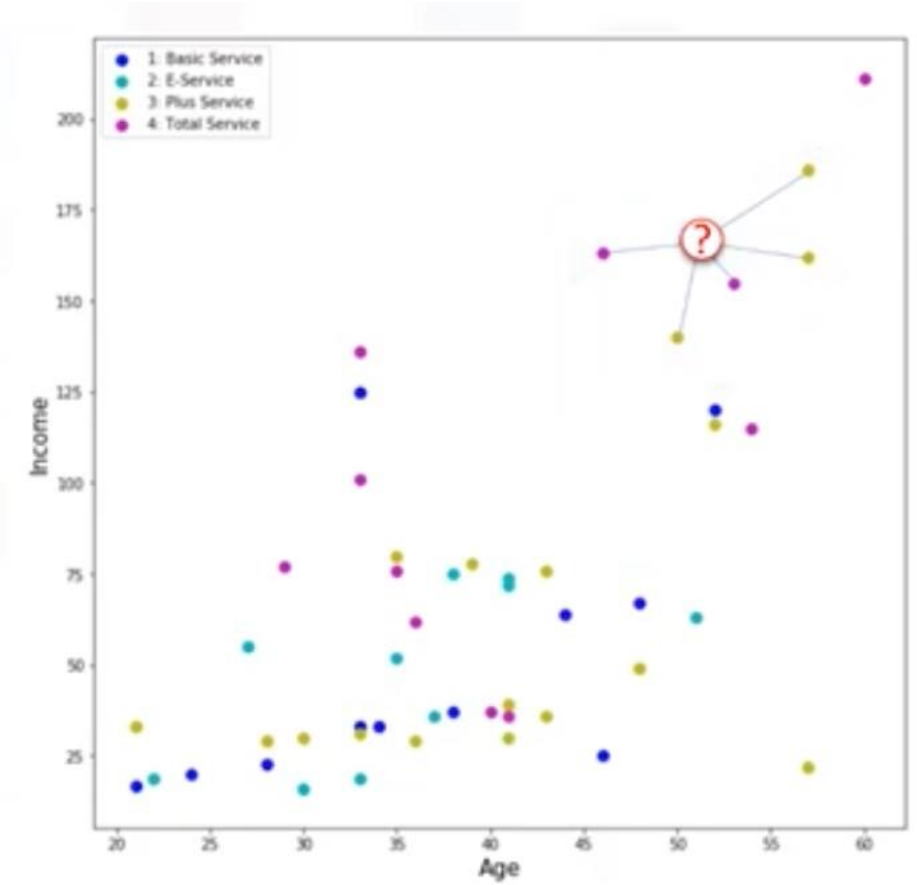
	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	13	44	1	9	64.0	4	5	0.0	0	2	1
1	3	11	33	1	7	136.0	5	5	0.0	0	6	4
2	3	68	52	1	24	116.0	1	29	0.0	1	2	3
3	2	33	33	0	12	33.0	2	0	0.0	1	1	1
4	2	23	30	1	9	30.0	1	2	0.0	0	4	3
5	2	41	39	0	17	78.0	2	16	0.0	1	1	3
6	3	45	22	1	2	19.0	2	4	0.0	1	5	2
7	2	38	35	0	5	76.0	2	10	0.0	0	3	4
8	3	45	59	1	7	166.0	4	31	0.0	0	5	3

Classification

KNN

- pick K
- calculate of the unknown points distance from all cases
- predict based on the K nearest points

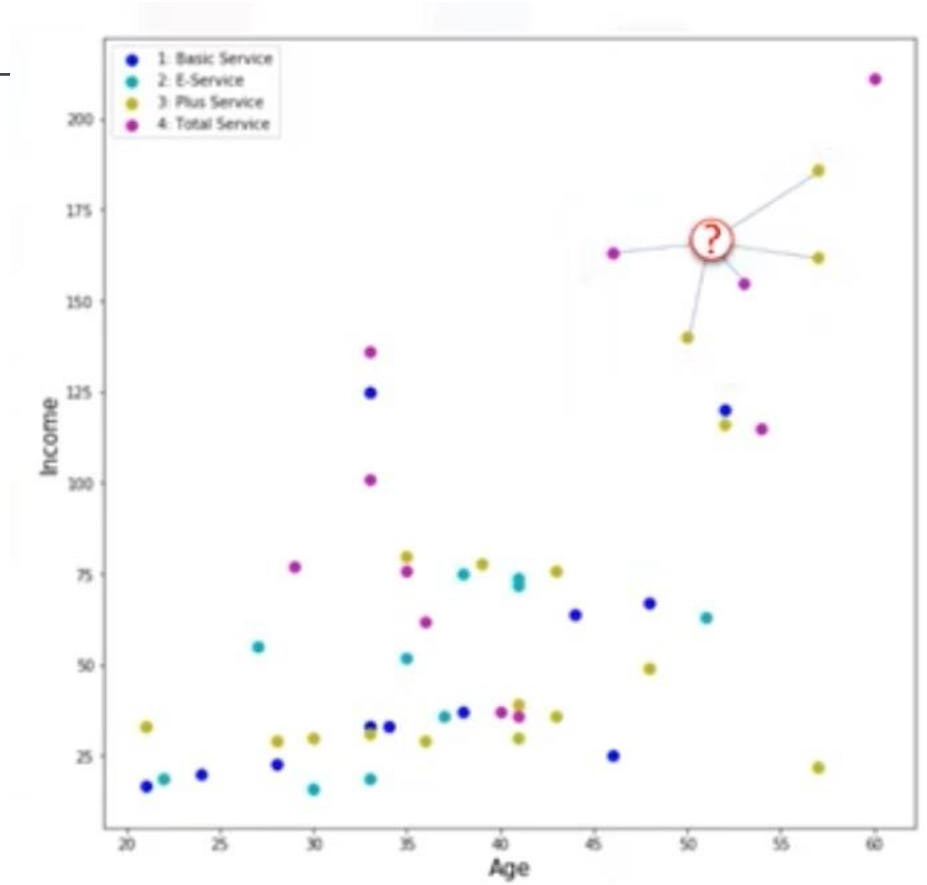
- How to find "distance" (Euclidean can be one way)
- How to choose K (low -> noise & overfit; high -> too general). Use the different Ks with test set and see which K is good.



Classification

KNN

- KNN can be used to compute a continuous target (regression)
- Say find 3 of the closest cases and find the median



Classification

KNN Evaluation

- Evaluation explains the performance of our model
- On test data we have y and $\hat{y}$
- There are different model evaluation metrics: Jaccard index, F1-score, and Log Loss.

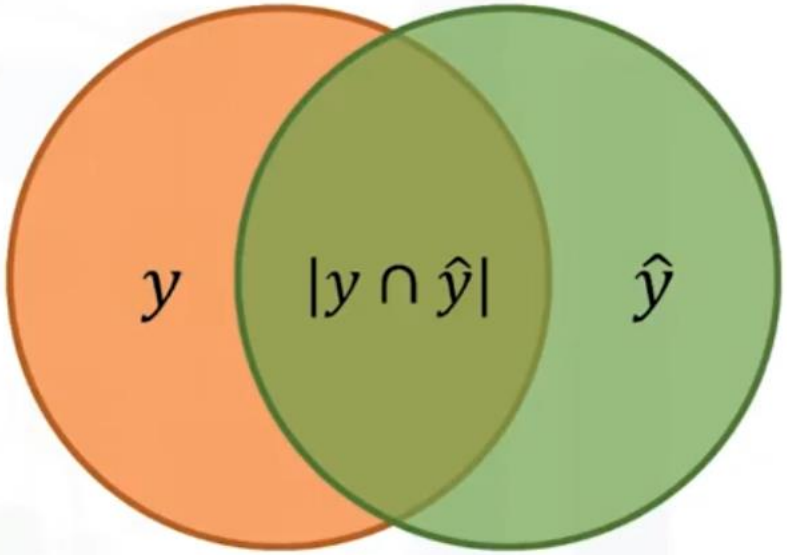
Classification

KNN Evaluation / Jaccard Index

y: Actual labels

$\hat{y}$: Predicted labels

$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$



y: [0, 0, 0, 0, 0, 1, 1, 1, ,1 ,1]

$\hat{y}$: [1, 1, 0, 0, 0, 1, 1, 1, ,1 ,1]

$$J(y,\hat{y}) = \frac{8}{10+10-8} = 0.66$$

Classification

KNN Evaluation / F1-Score

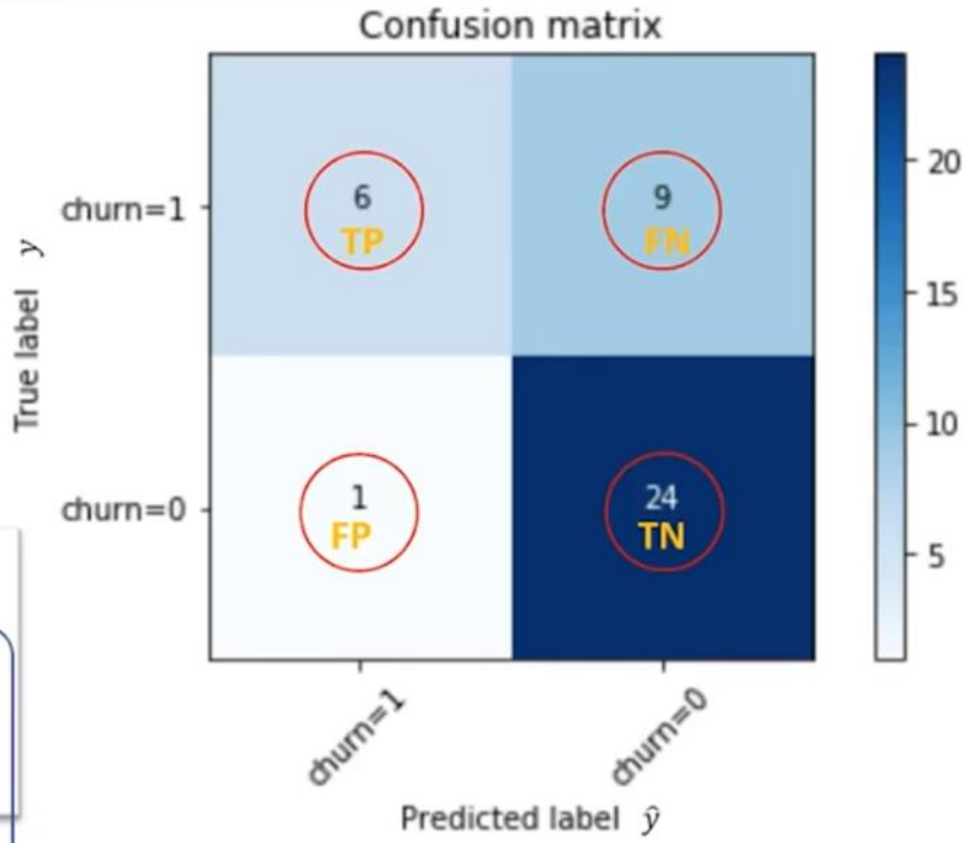
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-score = 2x (prc x rec) / (prc+rec)

F1-score: 0.00 ... 0.20 .... 0.55 .... 0.83 ... 1.00

Higher Accuracy

	precision	recall	f1-score
Churn = 0	0.73	0.96	0.83
Churn = 1	0.86	0.40	0.55

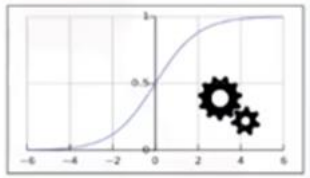
Avg Accuracy = 0.72



Classification

KNN Evaluation / LogLoss

Performance of a classifier where the predicted output is a probability value between 0 and 1.



Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

Test

Predicted churn	LogLoss
0.91	0.11
0.13	2.04
0.04	0.04
0.23	0.26
0.43	0.56

$LogLoss = 0.60$

Actual Labels y

$\hat{y}$ Predicted Probability

$$LogLoss = -\frac{1}{n}\sum(y \times \log(\hat{y}) + (1-y) \times \log(1-\hat{y}))$$

LogLoss: 0.00 ... 0.35 .... 0.60 ... 1.00

Higher Accuracy

Lab: KNN

- L04-MehrnazJalili