GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 11: NAIVE BAYESIAN CLASSIFICATION

January 12, 2015

DAT11-SF

LAST TIME:

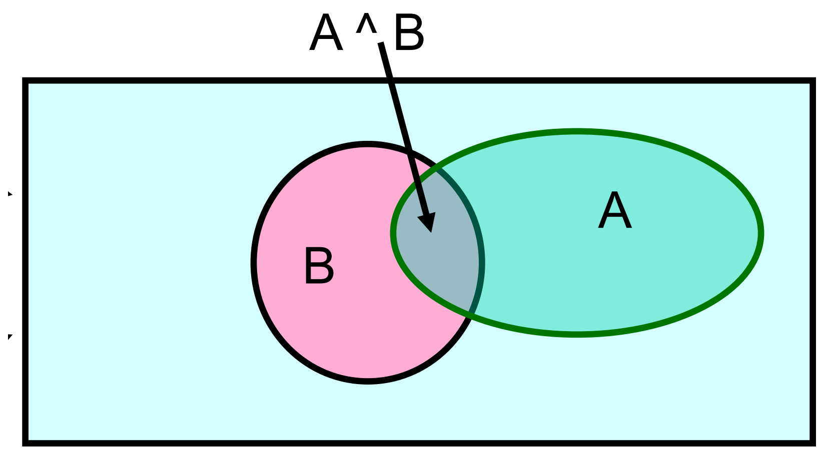
- ENSEMBLE METHODS
- BOOSTING: GRADIENT BOOSTING TREES
- BAGGING: RANDOM FOREST

QUESTIONS?

I. INTRO TO PROBABILITY
II. NAÏVE BAYESIAN CLASSIFICATION
III. RANKING CLASSIFIERS AND ROC CURVES

EXERCISES:
IV. NAÏVE BAYES IN SCIKIT-LEARN

I. INTRO TO PROBABILITY

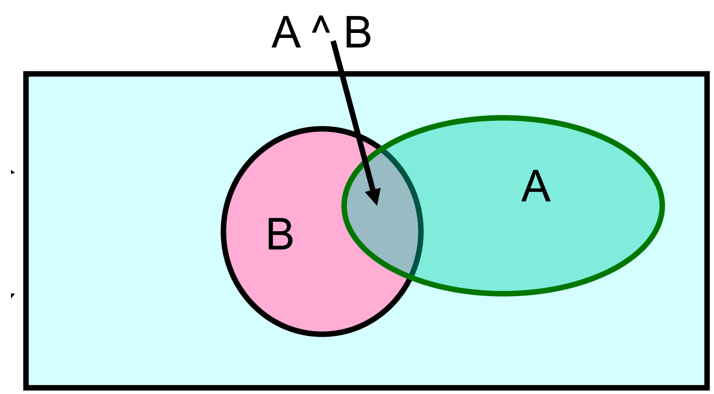


The probability of event A is denoted $P(A)$.

The probability of event B is denoted $P(B)$.

The probability of event $A \wedge B$ is denoted $P(AB)$.

Suppose event B has occurred. The probability of A **given** this information about B is called the **conditional probability** of A given B, written $P(A|B) = P(AB) / P(B)$.



The intersection of $A \wedge B$ divided by region B.

Q: What does it mean for two events to be **independent**?
A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

Using the definition of the conditional probability, we can also write:

$$P(A|B) = P(AB) / P(B) = P(A) \rightarrow P(AB) = P(A) * P(B)$$

$P(A|B) = P(AB) / P(B) \Rightarrow P(AB) = P(A|B)\,P(B)$

$P(AB) = P(A|B)*P(B)$

$P(BA) = P(B|A)*P(A)$

But $P(AB) = P(BA)$ since event AB = event BA

→ $P(A|B)*P(B) = P(B|A)*P(A)$ by combining the above

→ $P(A|B) = P(B|A)*P(A) / P(B)$ by rearranging last step

This result is called **Bayes' theorem**:

$P(A|B) = P(B|A) * P(A) / P(B)$

Thomas Bayes, 1701-1761

II. NAÏVE BAYESIAN CLASSIFICATION

h – hypothesis, D – data

P(h) – prior probability (apriory) of h

P(D|h) – probability of observing data under hypothesis h

P(h|D) – posterior probability of hypothesis h after seeing D

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Learning – selecting most probable hypothesis h after seeing data

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Select maximum a posteriory hypothesis:

$$\begin{aligned} h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}}\, P(h|D) \\ &= \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}}\, P(D|h)P(h) \end{aligned}$$

Suppose we have a dataset with features $x_1, \ldots, x_n$ and a class label C. What can we say about classification using Bayes' theorem?

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Bayes' theorem can help us to determine the probability of a new data point (record) belonging to a class, *given* the data we observe.

source: Data Analysis with Open Source Tools, by Philipp K. Janert. O'Reilly Media, 2011.

This term is the **likelihood function**. It represents the joint probability of observing features $\{x_i\}$ given that that record belongs to class C.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

We can observe the value of the likelihood function from the training data.

This term is the **prior probability** of C. It represents the probability of a record belonging to class C before the data is taken into account.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The value of the prior is also observed from the data.

This term is the **normalization constant.** It doesn't depend on C, and is generally ignored until the end of the computation.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The normalization constant doesn't tell us much.

This term is the **posterior probability** of C. It represents the probability of a record belonging to class C after the data is taken into account.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The goal of any Bayesian computation is to find ("learn") the posterior distribution of a particular variable.

Remember the likelihood function?

$$P(\{x_i\}\,|\,C) = P(\{x_1, x_2, \ldots, x_n\})\,|\,C)$$

Observing this exactly would require us to have enough data for every possible combination of features to make a reasonable estimate.

Make a simplifying assumption. In particular, we assume that the features x_i are conditionally independent from each other:

$$P(\{x_i\}|C) = P(x_1, x_2, \ldots, x_n)|C) \approx P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C)$$

This “naïve” assumption simplifies the likelihood function to make it tractable.

Selecting most probable MAP (maximum a-posteriory) hypothesis:

Assign class labels $\hat{y} = C_k$ according to :

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \ p(C_k) \prod_{i=1}^{n} p(x_i | C_k).$$

- Gaussian NB – continuous data

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

- Multinomal NB – discrete counts (histogram of event counts)

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

- Bernoully NB – discrete counts, boolean/binary variables

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{n} \left(x_i p_{ki} + (1 - x_i)(1 - p_{ki})\right)$$

TITANIC DATA

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoorte	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johans	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S

Find $P(Class = Yes \mid Status{=}First, Age = Adult, Sex = Male)$?

ATTRIBUTE	VALUE	CLASS=YES	CLASS=NO
STATUS	FIRST	203	122
	SECOND	118	167
	THIRD	178	528
	CREW	212	673
AGE	ADULT	654	1438
	CHILD	57	52
SEX	MALE	367	1364
	FEMALE	344	126

$$P(Status{=}First, Age = Adult, Sex = Male \mid Class = Yes) * P(Class = Yes) =$$

$$= P(Status{=}First \mid Class{=}Yes) * P(Age{=}Adult \mid Class{=}Yes) * P(Sex{=}Male \mid Class{=}Yes) * P(Class = Yes)$$

INTRO TO DATA SCIENCE

III. RANKING CLASSIFIERS AND ROC CURVES

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

(a) Confusion Matrix

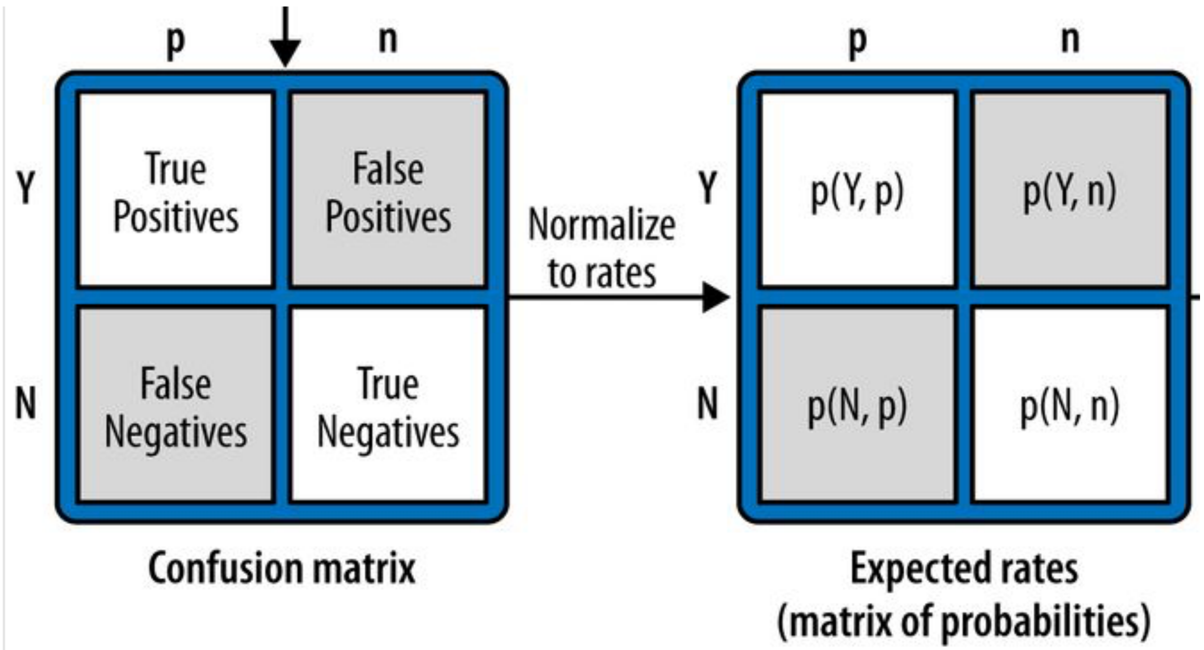
$$\text{Recall} = \frac{TP}{TP+FN}$$

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

$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

(b) Definitions of metrics



$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

Accuracy is a property of a classifier + data

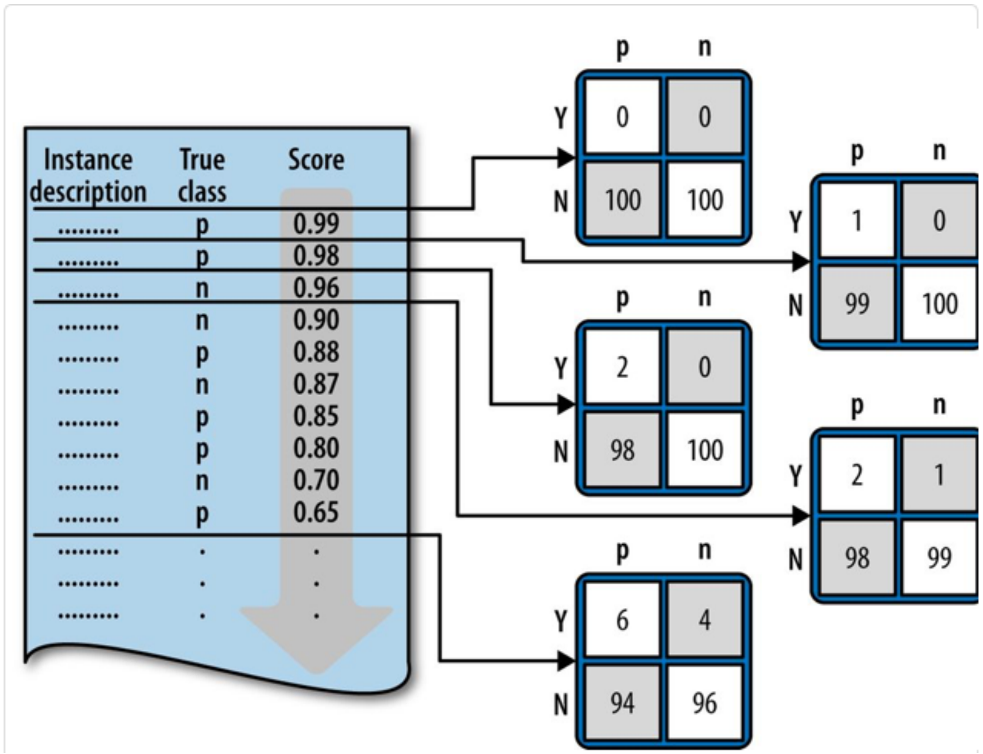
$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

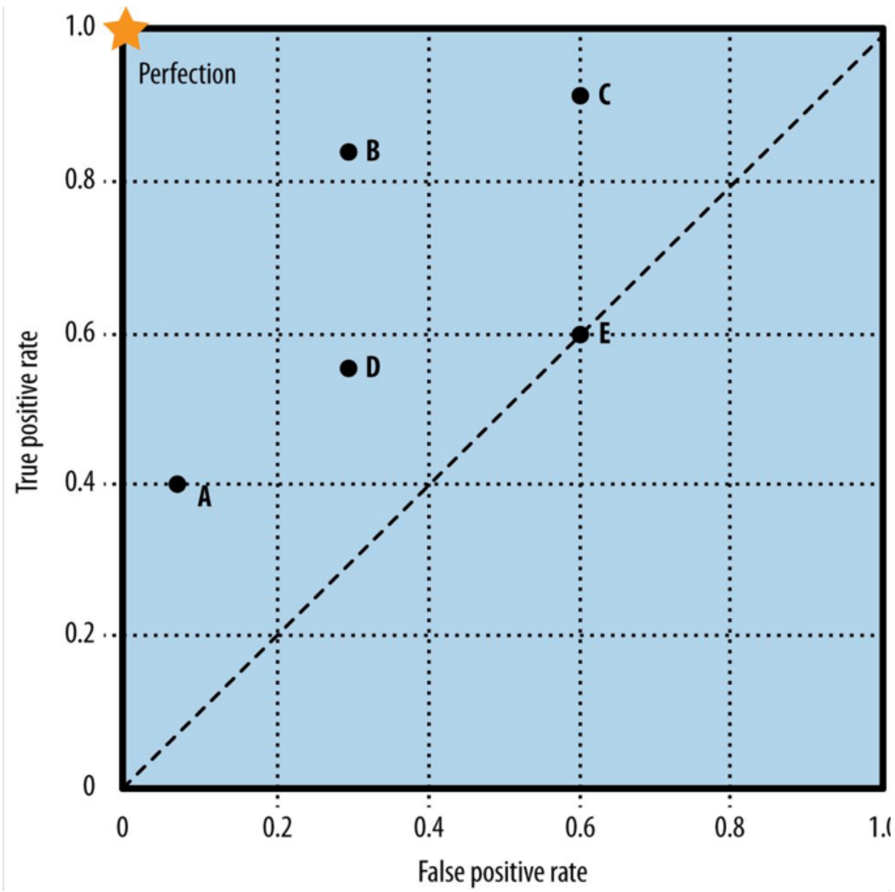
$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

True class

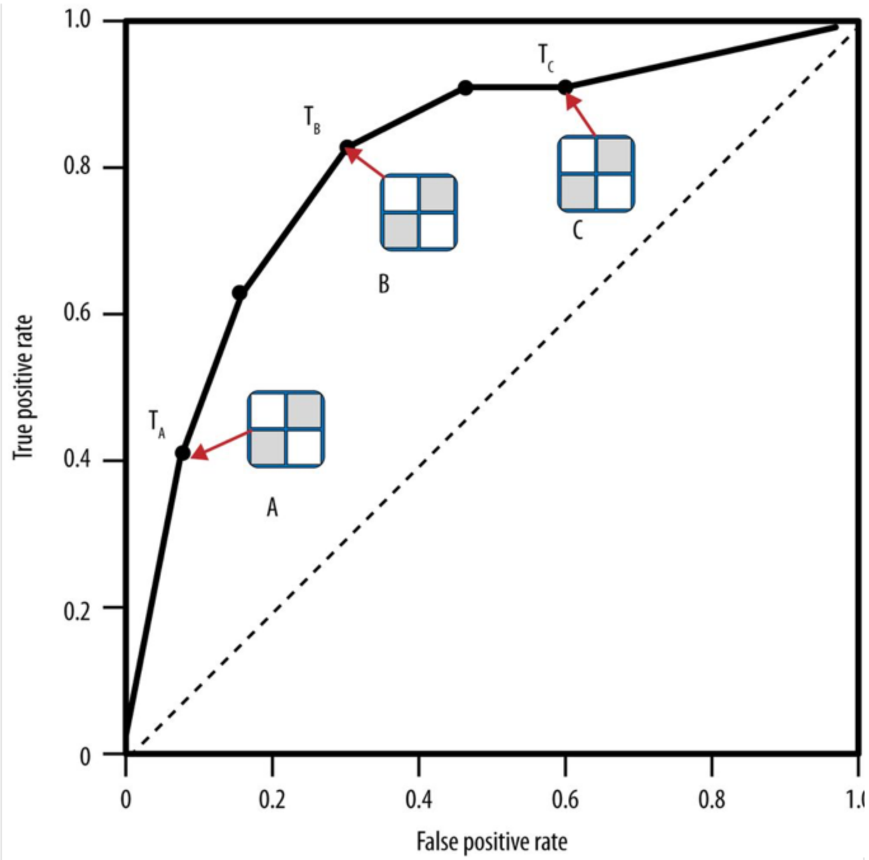
	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives
ls:	P	N


Instance description
True class
Score
.... p 0.99
.... p 0.98
.... n 0.96
.... n 0.90
.... p 0.88
.... n 0.87
.... p 0.85
.... p 0.80
.... n 0.70
.... p 0.65
p n
Y 0 0
N 100 100
p n
Y 1 0
N 99 100
p n
Y 2 0
N 98 100
p n
Y 2 1
N 98 99
p n
Y 6 4
N 94 96

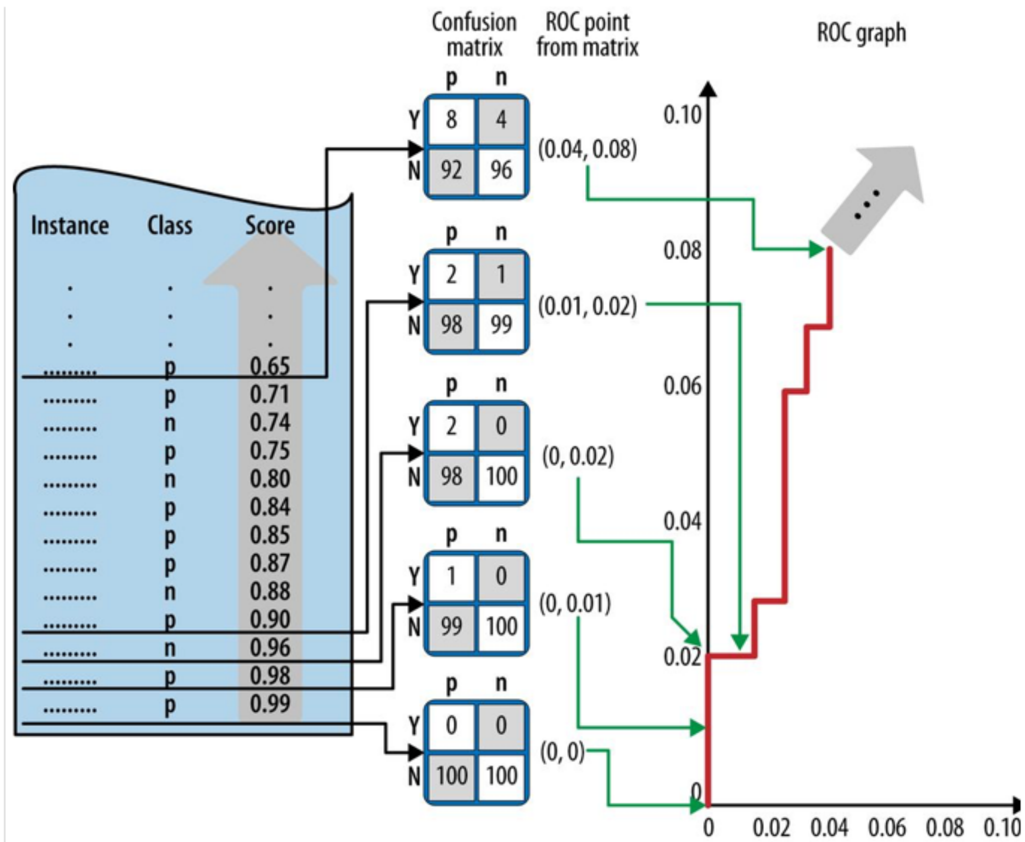


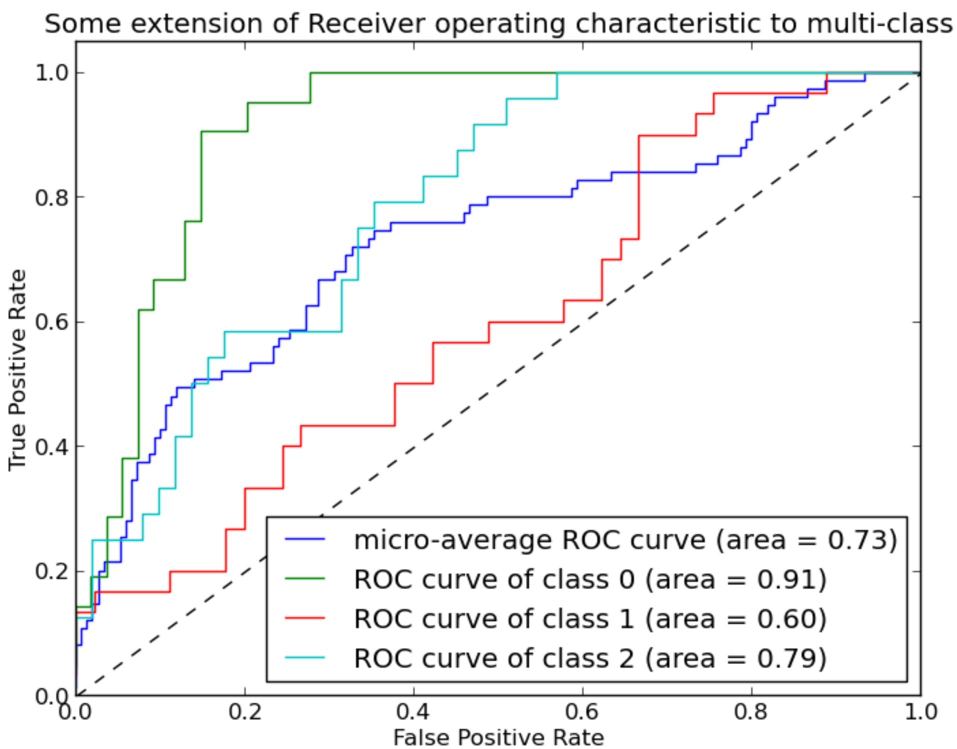
$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$



ROC =Receiver Operating Characteristic


Confusion matrix
ROC point from matrix
ROC graph
Instance
Class
Score
p 0.65
p 0.71
n 0.74
p 0.75
n 0.80
p 0.84
p 0.85
p 0.87
n 0.88
p 0.90
n 0.96
p 0.98
p 0.99
p n
Y 8 4
N 92 96
(0.04, 0.08)
Y 2 1
N 98 99
(0.01, 0.02)
Y 2 0
N 98 100
(0, 0.02)
Y 1 0
N 99 100
(0, 0.01)
Y 0 0
N 100 100
(0, 0)
0.10
0.08
0.06
0.04
0.02
0
0 0.02 0.04 0.06 0.08 0.10



AUC =Area under the ROC curve

INTRO TO DATA SCIENCE

III. LAB: NAÏVE BAYES IN SCI-KIT LEARN