10-601
Machine Learning

Naïve Bayes classifiers

Types of classifiers

- We can divide the large variety of classification approaches into three major types

1. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

2. Generative:
   - build a generative statistical model
   - e.g., Bayesian networks

3. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., decision tree

Bayes decision rule

- If we know the conditional probability P(X | y) we can determine the appropriate class by using Bayes rule:

$$P(y=i \mid X) = \frac{P(X \mid y=i)P(y=i)}{P(X)} \stackrel{def}{=} q_i(X)$$

But how do we determine p(X|y)?

Computing p(X|y)

Recall…

y – the class label

X – input attributes (features)

- Consider a dataset with 16 attributes (lets assume they are all binary). How many parameters to we need to estimate to fully determine p(X|y)?

age	employme	education	edun	marital	…	job	relation	race	gender	hour	country	wealth
					…							
39	State_gov	Bachelors	13	Never_mar	…	Adm_cleric	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	…	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	…	Handlers_c	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	…	Handlers_c	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	…	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	…	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	…	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	…	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	…	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	…	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	…	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	…	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	…	Adm_cleric	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acc	12	Never_mar	…	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	…	Craft_repai	Husband	Asian	Male	40	*MissingVa	rich
34	Private	7th_8th	4	Married	…	Transport_	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	…	Farming_fi	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	…	Machine_o	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	…	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	…	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	…	Prof_speci	Husband	White	Male	60	United_Sta	rich

Learning the values for the full conditional probability table would require enormous amounts of data

Naïve Bayes Classifier

• Naïve Bayes classifiers assume that given the class label (Y) the attributes are **conditionally independent** of each other:

$$p(X \mid y) = \prod_j p_j(x^j \mid y)$$

Product of probability terms

Specific model for attribute *j*

• Using this idea the full classification rule becomes:

$$\begin{aligned}\hat{y} &= \arg\max_v \; p(y = v \mid X) \\ &= \arg\max_v \frac{p(X \mid y = v)\, p(y = v)}{p(X)} \\ &= \arg\max_v \prod_j p_j(x^j \mid y = v)\, p(y = v)\end{aligned}$$

v are the classes we have

Conditional likelihood: Full version

$$L(X_i \mid y_i = 1, \Theta) = \prod_j p(x_i^j \mid y_i = 1, \theta_1^j)$$

Vector of binary attributes for sample *i*

The set of all parameters in the NB model

The specific parameters for attribute *j* in class 1

Note the following:

1. We assumes conditional independence between attributes **given** the class label
2. We learn a **different** set of parameters for the two classes (class 1 and class 2).

Learning parameters

$$L(X_i \mid y_i = 1, \Theta) = \prod_j p(x_i^j \mid y_i = 1, \theta_1^j)$$

- Let $X_1 \ldots X_{k1}$ be the set of input samples with label 'y=1'
- Assume all attributes are **binary**
- To determine the MLE parameters for $p(x^j = 1 \mid y = 1)$

  we simply count how many times the j'th entry of those samples in class 1 is 0 (call it n0) and how many times its 1 (n1). Then we set:

$$p(x^j = 1 \mid y = 1) = \frac{n1}{n0 + n1}$$

Final classification

- Once we computed all parameters for attributes in both classes we can easily decide on the label of a **new** sample X.

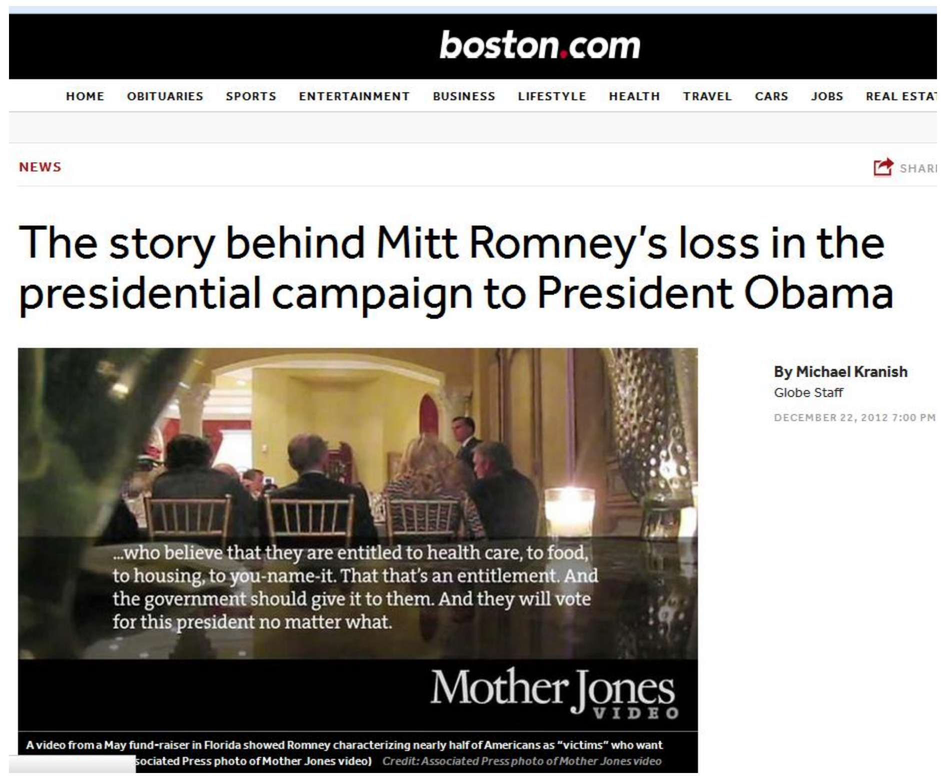
$$
\begin{aligned}
\hat{y} &= \arg\max_v \, p(y = v \mid X) \\
&= \arg\max_v \, \frac{p(X \mid y = v)\, p(y = v)}{p(X)} \\
&= \arg\max_v \prod_j p_j(x^j \mid y = v)\, p(y = v)
\end{aligned}
$$

Perform this computation for both class 1 and class 2 and select the class that leads to a higher probability as your decision

Prior on the prevalence of samples from each class

Example: Text classification

- What is the major topic of this article?


boston.com

HOME OBITUARIES SPORTS ENTERTAINMENT BUSINESS LIFESTYLE HEALTH TRAVEL CARS JOBS REAL ESTAT

NEWS SHAR

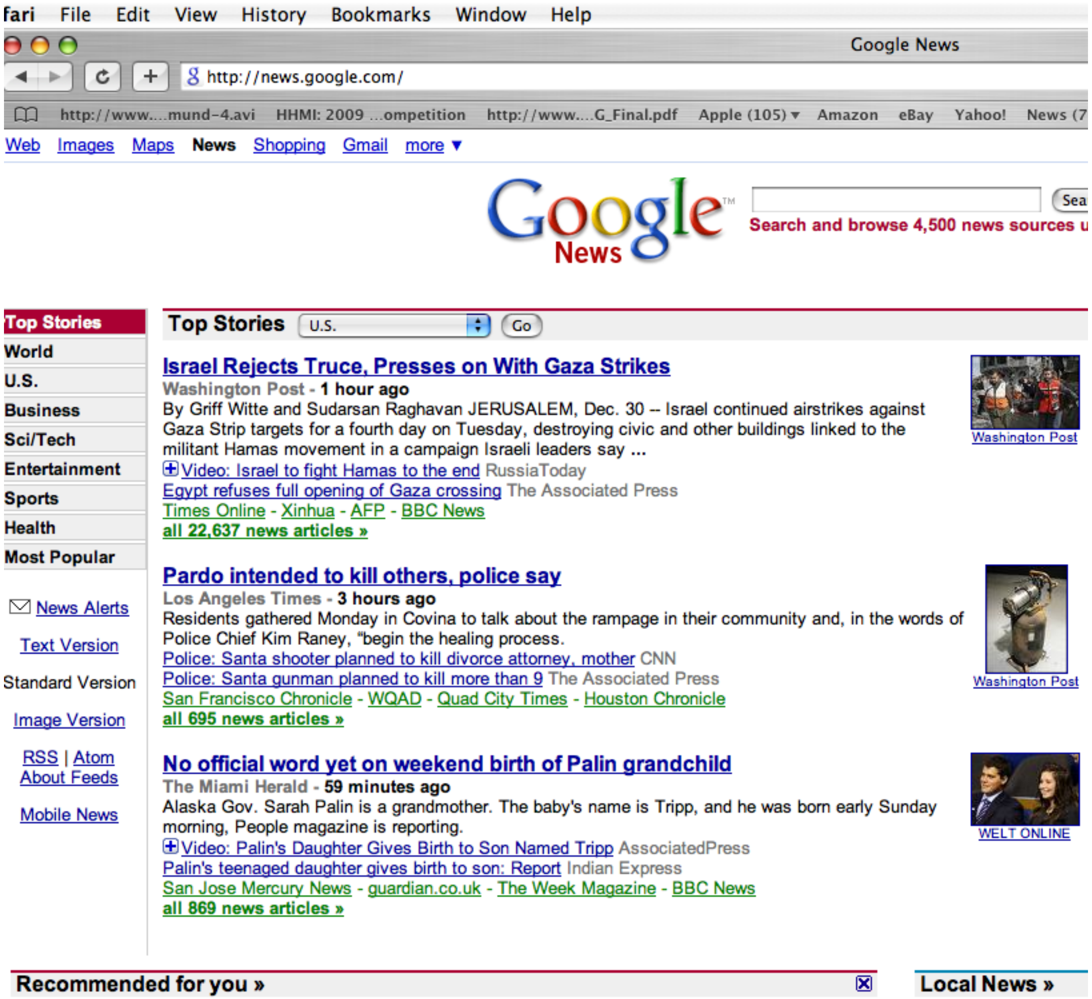
The story behind Mitt Romney's loss in the presidential campaign to President Obama

A video from a May fund-raiser in Florida showed Romney characterizing nearly half of Americans as "victims" who want
sociated Press photo of Mother Jones video) Credit: Associated Press photo of Mother Jones video

By Michael Kranish
Globe Staff
DECEMBER 22, 2012 7:00 PM

Example: Text classification

- Text classification is all around us



Feature transformation

- How do we encode the set of features (words) in the document?
- What type of information do we wish to represent? What can we ignore?
- Most common encoding: '**Bag of Words**'
- Treat document as a collection of words and encode each document as a vector based on some dictionary
- The vector can either be binary (present / absent information for each word) or discrete (number of appearances)

- Google is a good example
- Other applications include job search adds, spam filtering and many more.

Feature transformation: Bag of Words

- In this example we will use a binary vector
- For document X_i we will use a vector of *m** indicator features $\{\phi^j(X_i)\}$ for whether a word appears in the document
  - $\phi^j(X_i) = 1$, if word *i* appears in document x_j;
    $\phi^j(X_i) = 0$ if it does not appear in the document
- $\Phi(X_i) = [\phi^1(X_i) \ldots \phi^m(X_i)]^T$ is the resulting feature vector for the entire dictionary
- For notational simplicity we will replace each document X_i with a fixed length vector $\Phi_j = [\phi^1 \ldots \phi^m]^T$, where $\phi^j = \phi^j(X_i)$.

*The size of the vector for English is usually ~10000 words

Example

Assume we would like to classify documents as election related or not.

Dictionary

- Washington
- Congress

…

54. Romney
55. Obama
56. Nader


boston.com

HOME OBITUARIES SPORTS ENTERTAINMENT BUSINESS LIFESTYLE HEALTH TRAVEL CARS JOBS REAL ESTA

NEWS

SHAR

The story behind Mitt Romney's loss in the presidential campaign to President Obama

By Michael Kranish
Globe Staff
DECEMBER 22, 2012 7:00 PM

A video from a May fund-raiser in Florida showed Romney characterizing nearly half of Americans as "victims" who want
sociated Press photo of Mother Jones video) Credit: Associated Press photo of Mother Jones video

$$\phi^{54}=\phi^{54}(X_i) = 1$$
$$\phi^{55}=\phi^{55}(X_i) = 1$$
$$\phi^{56}=\phi^{56}(X_i) = 0$$

Example: cont.

We would like to classify documents as election related or not.

- Given a collection of documents with their labels (usually termed 'training data') we learn the parameters for our model.
- For example, if we see the word 'Obama' in *n1* out of the *n* documents labeled as 'election' we set *p('obama'|'election')=n1/n*
- Similarly we compute the priors (*p('election')*) based on the proportion of the documents from both classes.

Example: Classifying Election (E) or Sports (S)

Assume we learned the following model

$P(\phi^{romney}=1 \mid E) = 0.8$, $P(\phi^{romney}=1 \mid S) = 0.1$ $P(S) = 0.5$

$P(\phi^{obama}=1 \mid E) = 0.9$, $P(\phi^{obama}=1 \mid S) = 0.05$ $P(E) = 0.5$

$P(\phi^{clinton}=1 \mid E) = 0.9$, $P(\phi^{clinton}=1 \mid S) = 0.05$

$P(\phi^{football}=1 \mid E) = 0.1$, $P(\phi^{football}=1 \mid S) = 0.7$

For a specific document we have the following feature vector

$\phi^{romney} = 1$ $\phi^{obama} = 1$ $\phi^{clinton} = 1$ $\phi^{football} = 0$

$P(y = E \mid 1,1,1,0) \propto 0.8*0.9*0.9*0.9*0.5 = 0.5832$

$P(y = S \mid 1,1,1,0) \propto 0.1*0.05*0.05*0.3*0.5 = 0.000075$

So the document is classified as ‘Election’

Naïve Bayes classifiers for continuous values

- So far we assumed a binomial or discrete distribution for the data given the model ($p(X_i|y)$)
- However, in many cases the data contains continuous features:
  - Height, weight
  - Levels of genes in cells
  - Brain activity
- For these types of data we often use a Gaussian model
- In this model we assume that the observed input vector X is generated from the following distribution

$$X \sim N(\mu, \Sigma)$$

Gaussian Bayes Classification

- To determine the class when using the Gaussian assumption we need to compute p(X|y):

$$P(y = v \mid X) = \frac{p(X \mid y = v) P(y = v)}{p(X)}$$

$$P(X \mid y) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left[-\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right]$$

Once again, we need lots of data to compute the values of the mean μ and the covariance matrix Σ

Gaussian Bayes Classification

- Here we can also use the Naïve Bayes assumption: Attributes are independent given the class label
- In the Gaussian model this means that the covariance matrix becomes a **diagonal matrix** with zeros everywhere except for the diagonal
- Thus, we only need to learn the values for the variance term for each attribute: $x^j \sim N(\mu^j, \sigma^j)$

$$P(X \mid y = v) = \prod_j \frac{1}{(2\pi)^{1/2}\sigma_v^j} \exp\left[-\frac{\left(\mathbf{x}_j - \mu_v^j\right)^2}{2(\sigma_v^j)^2}\right]$$

Separate means and variance for each class

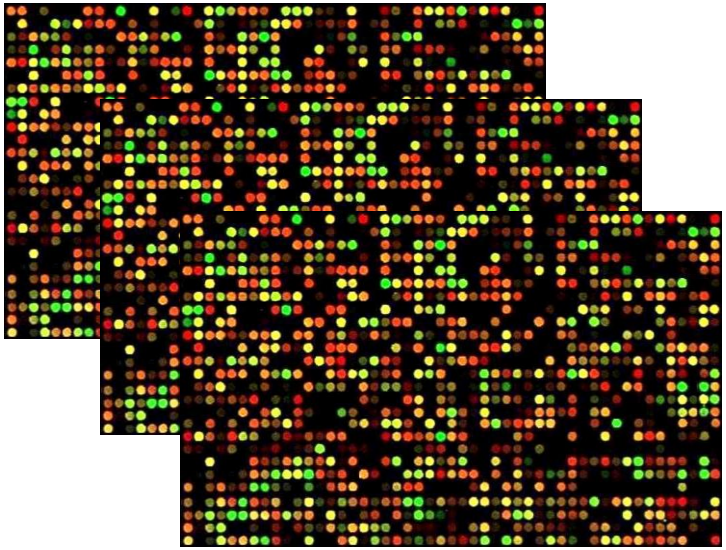
MLE for Gaussian Naïve Bayes Classifier

- For each class we need to estimate one global value (prior) and two values for each feature (mean and variance)
- The prior is computed in the same way we did before (counting) which is the MLE estimate For each feature
- Let the numbers of input samples in class 1 be k1. The MLE for mean and variance is computed by setting:

$$\mu_1^j = \frac{1}{k1} \sum_{X_i | s.t. y_i = 1} x_i^j \qquad \sigma_1^{j\,2} = \frac{1}{k1} \sum_{X_i | s.t. y_i = 1} (x_i^j - \mu_1^j)^2$$

Example: Classifying gene expression data

- Measures the levels (up or down) of genes in our cells
- Differs between healthy and sick people and between different disease types
- Given measurement of patients with two different types of cancer we would like to generate a classifier to distinguish between them



Classifying cancer types

- We select a subset of the genes (more in our 'feature selection' class later in the course).
- We compute the mean and variance for each of the genes in each of the classes
- Compute the class priors based on the input samples

Class 1 (ALL)

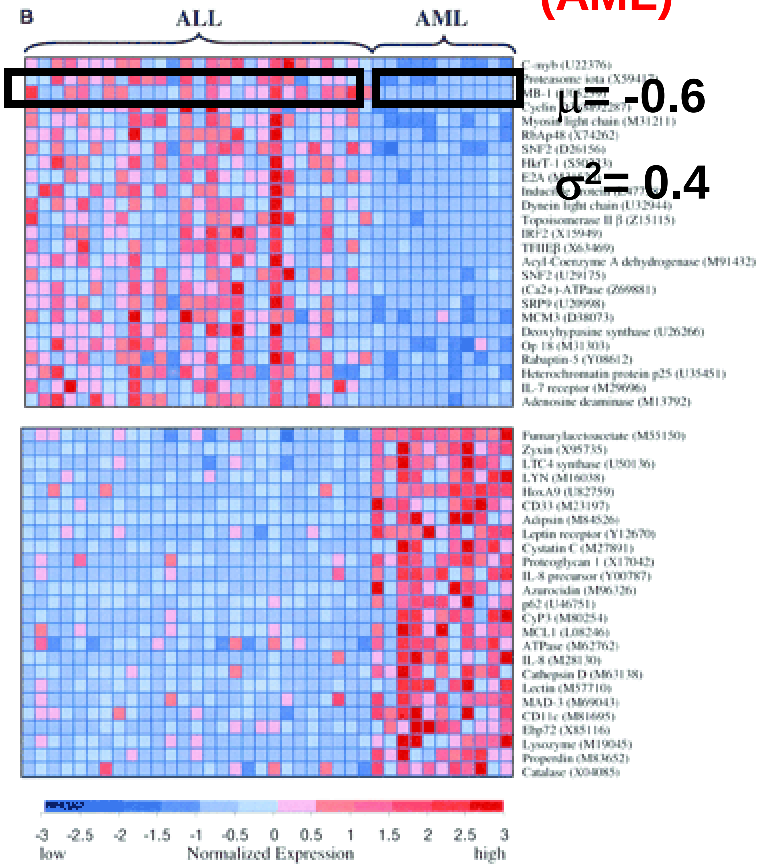
$\mu = 1.8$

$\sigma^2 = 1.1$

Class 2 (AML)

$\mu = -0.6$

$\sigma^2 = 0.4$



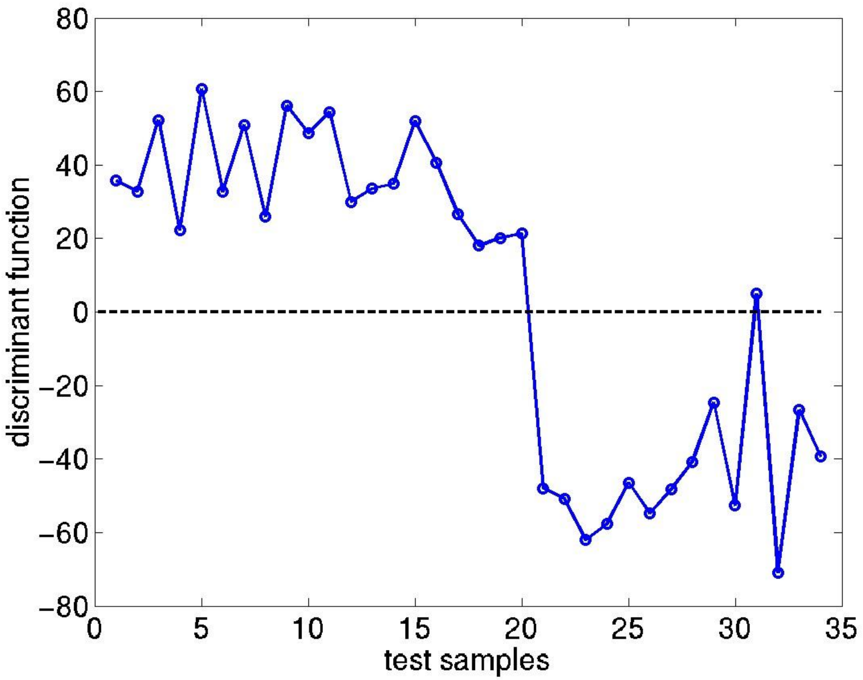
Classification accuracy

- The figure shows the value of the discriminate function

$$f(x) = \log \frac{p(y=1 \mid X)}{p(y=0 \mid X)}$$

across the test examples

- The only test error is also the decision with the lowest confidence



FDA Approves Gene-Based Breast Cancer Test*

“ MammaPrint is a DNA microarray-based test that measures the activity of 70 genes... The test measures each of these genes in a sample of a woman's breast-cancer tumor and then uses a specific formula to determine whether the patient is deemed low risk or high risk for the spread of the cancer to another site.”

*Washington Post, 2/06/2007

Possible problems with Naïve Bayes classifiers: Assumptions

- In most cases, the assumption of conditional independence given the class label is violated

  - much more likely to find the word ‘Barack’ if we saw the word ‘Obama’ regardless of the class
- This is, unfortunately, a major shortcoming which makes these classifiers inferior in many real world applications (though not always)
- There are models that can improve upon this assumption without using the full conditional model (one such model are Bayesian networks which we will discuss later in this class).

Possible problems with Naïve Bayes classifiers: Parameter estimation

- Even though we need far less data than the full Bayes model, there may be cases when the data we have is not enough
- For example, what is $p(S=1,N=1|E=2)$?
- This can get worst. Assume we have 20 variables, almost all pointing in the direction of the same class except for one for which we have no record for this class.
- Solutions?

Summer?	Num > 20	Evaluation
1	1	3
1	0	3
0	1	2
0	1	1
0	0	3
1	1	1

Important points

- Problems with estimating full joints
- Advantages of Naïve Bayes assumptions
- Applications to discrete and continuous cases
- Problems with Naïve Bayes classifiers