COMP47750/COMP47990 Tutorial

Naïve Bayes Classifiers

1. A training set contains 10 examples describing weather conditions in terms of 5 categorical features. The classification task is to predict whether an individual will go swimming based on these conditions (i.e. the class labels are either "Yes" or "No").

Based on this training set, we are given the following contingency table of conditional and prior class probabilities:

Swimming	Yes	No
Rain Recently=light	0/4	3/6
Rain Recently=moderate	2/4	3/6
Rain Recently=heavy	2/4	0/6
Rain Today=light	1/4	3/6
Rain Today=moderate	2/4	3/6
Rain Today=heavy	1/4	0/6
Temp=Cold	1/4	5/6
Temp=Warm	3/4	1/6
Wind=Light	2/4	2/6
Wind=Moderate	2/4	2/6
Wind=Gale	0/4	2/6
Sunshine=Some	2/4	4/6
Sunshine=None	2/4	2/6
Class Probabilities	4/10	6/10

Using the contingency table above, classify the two new examples below using Naïve Bayes.

	Rain Recently (RR)	Rain Today (RT)	Temp (T)	Wind (W)	Sunshine (S)	Swimming
X1	Heavy	Moderate	Warm	Light	Some	???
X2	Light	Moderate	Warm	Light	Some	???

2. Consider the following dataset, which contains examples describing several cases of sunburn.

	Name	Hair	Height	Build	Lotion	Result
1	Sarah	blonde	average	light	no	sunburned
2	Dana	blonde	tall	average	yes	none
3	Alex	brown	short	average	yes	none
4	Annie	blonde	short	average	no	sunburned
5	Emily	red	average	heavy	no	sunburned
6	Pete	brown	tall	heavy	no	none
7	John	brown	average	heavy	no	none
8	Katie	brown	short	light	yes	none

a) Construct the contingency table that would be used by Naïve Bayes to build a classifier for this dataset.

b) Use Naïve Bayes to give the likelihood that the result for the given example X is "sunburned". Then indicate what prediction Naïve Bayes would make.

	Hair	Height	Build	Lotion	Result
X	blonde	average	heavy	no	???

3. Consider the following dataset in a task that aims to predict the risk of a loan application based on 3 features describing each applicant: credit history, debt, and income. Applications are assigned to 3 different risk classes: {low, medium, high}.

	Credit History	**Debt**	**Income**	**Risk**
1	bad	low	0to30	high
2	bad	high	30to60	high
3	bad	low	0to30	high
4	unknown	high	30to60	high
5	unknown	high	0to30	high
6	good	high	0to30	high
7	bad	low	over60	medium
8	unknown	low	30to60	medium
9	good	high	30to60	medium
10	unknown	low	over60	low
11	unknown	low	over60	low
12	good	low	over60	low
13	good	high	over60	low
14	good	high	over60	low

b) Construct a contingency table that would be used by Naïve Bayes to build a classifier using this training data.

b) Based on the contingency table, predict the risk level for the new loan application X below.

	Credit History	**Debt**	**Income**	**Risk**
X	bad	low	30to60	???

4. (a) Given the nature of the `AthleteSelection` data which would be the best of the Naive Bayes options in scikit-learn for that classification task?

(b) A ranking classifier is a classifier that can rank a test set in order of confidence for a given classification outcome. Naive Bayes is a ranking classifier because the ‘probability’ can be used as a confidence measure for ranking.

1. Train a Naive Bayes classifier from the `AthleteSelection` data. Load the test data from `AthleteTest.csv` and apply the classifier.
2. Use the `predict_proba` method to find the probability of being selected.
3. Rank the test set by probability of being selected.
   3.1. Who is most likely to be selected?
   3.2. Who is least likely?

Some code for this exercise is available in the notebook `07 Naive Bayes Tutorial`. You will also need to download the test data file ‘AthleteTest.csv’.

(c) When a **`GaussianNB`** model is trained the model is stored in two parameters **‘theta_’** and **‘var_’.** Train a **`GaussianNB`** model and check to see if these parameters agree with your own estimates.

Hint: this code will give you the estimates you need.

```
athlete[athlete[‘Selected']=='No']['Agility'].mean()
athlete[athlete[‘Selected’]=='No']['Agility'].var(ddof=0)
```

The **‘var_’** parameter contains the square of the standard deviation (the variance) rather than the standard deviations. The figures should agree exactly if **ddof** is set to zero is the **var** calculation.