Naïve Bayes Classification

Introduction

- The Bayes Classifier requires probability structure of the problem to be known.
- Density estimation (using non-parametric or parametric methods) is one way to handle the problem.
- There are several problems ….

Problems with density estimation

- Large datasets are needed.
- Numeric valued features are required.

- In practice these two may not be satisfied.

How to overcome the problem

- One has to work with the given data set.
- So, Probability estimations needs to be done using the given data only.
- Often marginal probabilities can be better estimated than the joint probabilities.
- Also, marginal probabilities are easy to compute.

Play-tennis data

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

- P(<sunny,cool,high,false>|N) = 0
- But, P(sunny|N) = 3/5, P(cool|N) = 1/5, P(high|N) = 4/5, P(false|N) = 2/5.

- P(<sunny, cool, high, false>|N) = 0
- This may be because of the smaller dataset.
- If we increase the dataset size, this may become a positive number.

- This problem is often referred to as “the curse of dimensionality”.

Assumption

- Make the assumption that for a given class, features are independent of each other.
- In practice, this assumption holds very often.

- Then P(<sunny,cool,high,false>|N) = P(sunny|N) . P(cool|N) . P(high|N) . P(false|N) = 3/5 . 1/5 . 4/5 . 2/5 = 24/625.

Naïve Bayesian Classification

- Naïve assumption: for a given class, features are independent of each other

$$P(<x_1,\ldots,x_k>|C) = P(x_1|C)\cdot\ldots\cdot P(x_k|C)$$

- $P(x_i|C)$ is estimated as the relative freq of samples having value x_i as i-th attribute in class C
- It often makes the problem a feasible and easy one to solve.

Play-tennis example: estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14
P(n) = 5/14

outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 2/5**
windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

Play-tennis example: classifying X

- An unseen sample X = <rain, hot, high, false>

- P(X|p)·P(p) =

  P(rain|p)·P(hot|p)·P(high|p)·P(false|p)·P(p) = 3/9·2/9·3/9·6/9·9/14 =

  0.010582

- P(X|n)·P(n) =

  P(rain|n)·P(hot|n)·P(high|n)·P(false|n)·P(n) = 2/5·2/5·4/5·2/5·5/14 =

  0.018286

- Sample X is classified in class n (don't play)

With Continuous features

- In order to use the Naïve Bayes classifier, the features has to be discretized appropriately (otherwise what happens?)

With Continuous features

- In order to use the Naïve Bayes classifier, the features has to be discretized appropriately (otherwise what happens?)
- Height = 4.234   will not occur anywhere in that column; but 4.213, 4.285 may be occuring. If you discretize (eg., rounding) then fequency ratio’s are meaningful.
- Clustering of feature values of a feature may be done to achieve a better discretization.