Naïve Bayes Classification and Regression

Statement for Audio and Video Learning Resources

Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is approximately 70-90% accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.

If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.

Contents

- Why use statistics
- Calculating probabilities
- Bayesian theorem –Naïve Bayes
- Missing values
- Numeric values
- Summary

Why use statistics

- Simple algorithms often work well in practice.
- Based on using ALL the attributes for prediction.
- Assumptions
  - All attributes are equally important
  - Attributes are independent
    - I.e. the values of attributes are NOT related
- Gives good result even though assumptions are often incorrect.
- While here we focus on classification, it can be used for regression too.

Observed probability from datasets

Count the number in the dataset.

If there are 100 instances in the dataset and 2 of them satisfy your condition, then the probability of that condition is 2 out of 100 (which is 1 in 50).

Probabilities are like fractions.

Example

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Cloudy	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Cloudy	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Cloudy	Mild	High	True	Yes
Cloudy	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

14 instances in total

9 for Play=Yes

5 for Play=No

Example: frequency table

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	true	3	3	9	5
cloudy	4	0	mild	4	2	normal	6	1	false	6	2		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	true	3/9	3/5	9/14	5/14
cloudy	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	false	6/9	2/5		
rainy	3/9	2/5	cool	3/9	1/5								

3 outlook = rainy with play=yes out of 9 instances of play=yes

Example: outlook = rainy

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Cloudy	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Cloudy	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Cloudy	Mild	High	True	Yes
Cloudy	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

5 outlook = rainy

3 outlook = rainy with play = Yes

2 outlook = rainy with play = No

Calculating class for new example

- A new day:
  - Outlook = sunny
  - Temperature = cool
  - Humidity = high
  - Windy = true
- Likelihood
  - For yes: 2/9 * 3/9 * 3/9 * 3/9 * 9/14 = 0.0053
  - For no: 3/5 * 1/5 * 4/5 * 3/5 * 5/14 = 0.0206
- Probability (normalise so that probabilities add up to 1)
  - yes: $\frac{0.0053}{0.0053+0.0206}$= 0.205
  - no: $\frac{0.0206}{0.0053+0.0206}$ = 0.795
- So class is NO

Example: frequency table (sunny, cool, high, true)

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	true	3	3	9	5
cloudy	4	0	mild	4	2	normal	6	1	false	6	2		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	true	3/9	3/5	9/14	5/14
cloudy	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	false	6/9	2/5		
rainy	3/9	2/5	cool	3/9	1/5								

Naïve Bayes' Theorem

- Naïve Bayes' theorem

$$p(H|E) = \frac{p(E|H) * p(H)}{p(E)}$$

- **Posterior probability P(H|E):** the probability of H given that evidence E has been observed.
- **Prior probability P(H):** the probability of hypothesis H.
- **Marginal probability p(E):** the probability of evidence E.
- **Likelihood probability p(E|H):** the probability of evidence given hypothesis is true

Naïve Bayes' Theorem for classification

- Calculate the probability of the class given an instance
- Instance = evidence E
- Class value = hypothesis H
- Evidence is split into independent parts (attributes of instance)
- So the posterior probability is

$$p(H|E) = \frac{p(E_1|H) * p(E_2|H) * \ldots * p(E_n|H) * p(H)}{p(E)}$$

What about this? Ignore it for now.

Applying Naïve Bayes' Theorem

A new day with evidence E:

A: Outlook = sunny; B: Temperature = cool; C: Humidity = high; D: Windy = true

$$p(yes|E) = \frac{p(A|y) * p(B|y) * p(C|y) * p(D|y) * p(y)}{p(E)} = \frac{\frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} * \frac{9}{14}}{p(E)} = \frac{0.0053}{p(E)}$$

$$p(no|E) = \frac{p(A|n) * p(B|n) * p(C|n) * p(D|n) * p(n)}{p(E)} = \frac{\frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} * \frac{5}{14}}{p(E)} = \frac{0.0206}{p(E)}$$

Applying Bayes' Theorem

Normalising

- $p(y|E) = \frac{\frac{0.0053}{p(E)}}{p(y|E)+p(n|E)} = \frac{\frac{0.0053}{p(E)}}{\frac{0.0053}{p(E)}+\frac{0.0206}{p(E)}} = \frac{0.0053}{0.0053+0.0206} = 0.205$

- $p(n|E) = \frac{\frac{0.0206}{p(E)}}{p(y|E)+p(n|E)} = \frac{\frac{0.0206}{p(E)}}{\frac{0.0053}{p(E)}+\frac{0.0206}{p(E)}} = \frac{0.0206}{0.0053+0.0206} = 0.795$

All p(E) cancel out! Phew!

Problem: missing combinations

If a certain combination of values does NOT occur in the training set, its probability is 0

- Since probabilities are multiplied, the final probability is 0.
- E.g. we have no instance of *outlook = cloudy* with *play = no*
- We cannot predict something we cannot see?

Solution: add 1 to every count (Laplace estimator)

2 + 1 = 2 ‘sunny’ and ‘yes’ plus 1

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	3	4	hot	3	3	high	4	5	true	4	4	9	5
cloudy	5	1	mild	5	3	normal	7	2	false	7	3		
rainy	4	3	cool	4	2								
sunny	3/12	4/8	hot	3/12	3/8	high	4/11	5/7	true	4/11	4/7	9/14	5/14
cloudy	5/12	1/8	mild	5/12	3/8	normal	7/11	2/7	false	7/11	3/7		
rainy	4/12	3/8	cool	4/12	2/8								

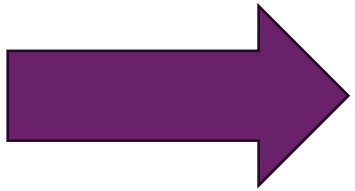
3+5+4 = number of ‘yes’ (9) plus number of values for outlook (3)

Better solution?

Laplace estimator: adding 1 to each count (why 1? No reason!)

- Instead, use a small constant **μ**.
- For example, the probabilities for outlook when play = yes change to

$\frac{2}{9}$	$\frac{2+\frac{\mu}{3}}{9+\mu}$
$\frac{4}{9}$	$\frac{4+\frac{\mu}{3}}{9+\mu}$
$\frac{3}{9}$	$\frac{3+\frac{\mu}{3}}{9+\mu}$



3 possible values

μ indicates how important the a priori value of the attribute is.

Better solution – add a "small" amount?

- There's no reason why all μ are multiplied by a 1/3 factor. These factors could be different for as long as they add up to 1.
  - $(2 + \mu * P_1) / (9 + \mu)$
  - $(4 + \mu * P_2) / (9 + \mu)$
  - $(3 + \mu * P_3) / (9 + \mu)$
- Advantage: rigorous. All a priori probabilities assigned.
- Disadvantage: how these probabilities **should** be assigned is unclear.
- In practice, the ***Laplace estimator*** is used.

Problem? Missing Values

If a new instance has an attribute which has a missing value, then it is omitted from the calculation.

- E.g. new instance with no outlook value, i.e.
  - Temperature = cool
  - Humidity = high
  - Windy = true
  - Likelihood(yes) = 3/9 * 3/9 * 3/9 * 9/14 = 0.0238
  - Likelihood(no) = 1/5 * 4/5 * 3/5 * 5/14 = 0.0343
  - Normalising
    - P(yes) = 0.0238/(0.0238 + 0.0343) = 0.41
    - P(no) = 0.0343/ (0.0238 + 0.0343) = 0.59

If a value is missing in a training instance, it is not included in the frequency count

Numeric values?

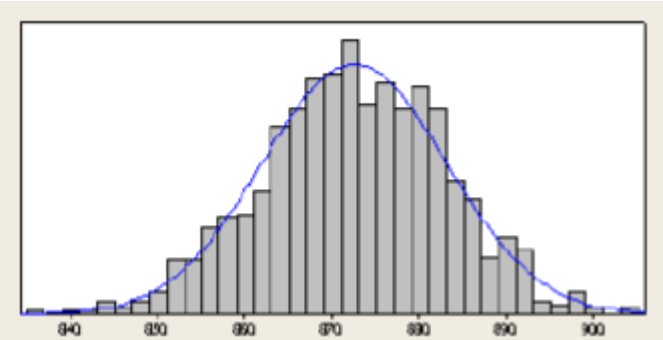
2 possibilities

- Either: Keep the values numeric, and use statistics about them
- Or: Discretise the numeric values
  - i.e. convert the numeric values to nominal ones
  - Sort examples according to numeric values for selected attribute, including class outcome.
  - Partition according to class category

Numeric values: keeping them numeric

Assume a Normal or Gaussian distribution

- Nominal values are calculated as before
  - Counts are normalized into probabilities
- Numeric values are listed
  - Instead of counts, calculate
    - Mean
    - Standard deviation



Numeric values:keeping them numeric

- Mean μ (average): add all the values together and divide by number of values.
- Standard deviation σ :
  - Subtract the mean from each value
  - Square the result
  - Add all results together
  - Divide by number of values – 1
  - Calculate square root

Example: Numeric values

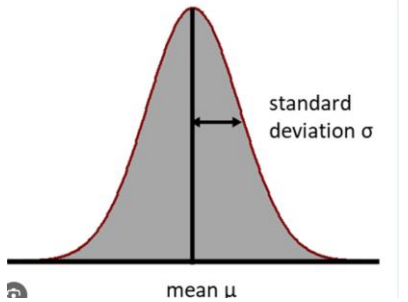
Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Cloudy	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Cloudy	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Cloudy	72	90	True	Yes
Cloudy	81	75	False	Yes
Rainy	71	91	True	no

Numeric values example

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83 70 68	85 80 65		86 96 80	85 90 70	true	3	3	9	5
cloudy	4	0		64 69 75	72 71		65 70 80	95 91	false	6	2		
rainy	3	2		75 72 81			70 90 75						
sunny	2/9	3/5	Mean μ	73	74.3	Mean μ	79.1	86.2	true	3/9	3/5	9/14	5/14
cloudy	4/9	0/5	Std dev σ	6.2	7.9	Std dev σ	10.2	9.7	false	6/9	3/5		
rainy	3/9	3/5											

Probability density function (Gaussian distrib.)

- Indicates the probability of a quantity being close to a given value



- $f(x) = \frac{1}{\sqrt{2\pi}*\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- For example, the probability of a temperature being *close* to 66
- $f(temperature = 66|yes) = \frac{1}{\sqrt{2\pi}*6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.034$

Numeric values – probability calculations

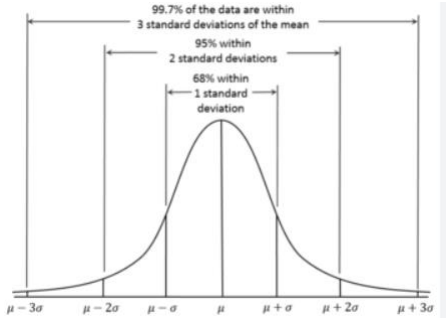
- New day:
  - outlook = sunny, temperature = 66, humidity = 90, windy = true.
- Likelihood (yes) =

  2/9 * 0.0340 * 0.0211 * 3/9 * 9/14 = 0.000036
- Likelihood(no) =

  3/5 * 0.0291 * 0.0380 * 3/5 * 5/14 = 0.000136
- Probability(yes) = 0.209
- Probability(no) = 0.791

Variants of Naïve Bayes

- The following variants are popular.
  - Gaussian (seen)
    - Assumes a normal distribution of (continuous) numeric attributes
  - Multinomial
    - Assumes **multinomial** distribution.
    - Used for text classification.
  - Bernoulli
    - Independent **binary** attributes.
    - Used for text classification.



The Binomial Distribution and Test, Clearly Explained!!! - Statquest

Advantages and Disadvantages of Naïve Bayes

- Advantages
  - Multi-class classification
  - Regression
  - Fast
  - Often used for text classification
- Disadvantages
  - Attribute independence.
  - All attributes are equally important.

Summary

- Probabilities can be used to predict new class or for regression.
- Simple method where normally assumptions are not correct, i.e. independent attributes, but it produces good results.
  - If several attributes are related (redundant attributes), their effect on the conclusion is magnified and Naïve Bayes' theorem does not work well.
  - If numeric values do not follow a normal distribution, the use of other formulas, suitable for that distribution, is required.
  - Discretisation of the numeric values is also a possibility.