Data Analysis and Knowledge Discovery

Bayes Classifiers

Jukka Heikkonen

University of Turku
Department of Information Technology

Jukka.Heikkonen@utu.fi

Naive Bayes

- classification method
- a probabilistic model, with very simple (unrealistically naive) independence assumptions
- pros: simple, scales well to large data sets, produces a very compact and fast model for prediction, can handle categorical variables naturally, robust to irrelevant features, can produce accurate models on simple enough problems, no hyperparameters to tune
- cons: might not produce accurate models on complex problems, not that great for handling numerical features

Naive Bayes and text classification

- typical application: text classification
- bag-of-words feature representation
  - document classification: does an e-mail belong to category JUNK or NON-JUNK
  - pre-processing: form a dictionary of all words appearing in training data
  - features: which words appear in a document
  - can be encoded as a feature vector, with each element of vector corresponding to a possible word
  - very long vectors filled with mostly zeroes (maybe 100 words out of 50000 possible appear in a given document...)
- How would you classify these instances, JUNK or NON-JUNK?:
  - {"and", "dear","deposit", "lifetime", "lottery", "opportunity", "or", "rich", "Sir/Madam"}
  - {"analysis", "and", "assignment", "course", "data", "next", "or", "week"...}

Naive Bayes

- we want a model that can estimate the following probability:
- given the features I have observed, what are the probabilities for this instance belonging to different classes
- Example: given the words in the e-mail, what is the probability of it belonging to class JUNK / NON-JUNK?

Bayes' theorem

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Interpretation: The probability that a hypothesis H is true given event E has occurred, can be derived from the probability

- of the hypothesis itself,
- of the event and
- the conditional probability of the event given the hypothesis.

Proof:

$$\frac{P(E|H) \cdot P(H)}{P(E)} = \frac{\frac{P(E \cap H)}{P(H)} \cdot P(H)}{P(E)} = P(H|E)$$

Bayes' theorem, junk-mail example

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

- H: "Is junk-mail"
- E: "Words observed in the e-mail"
- $P(H|E)$: "Given the words, how likely is this to be junk-mail?"
- $P(E|H)$: "Given that a message is junk-mail, how likely it is to contain these words?"
- $P(H)$: "How likely is a randomly selected e-mail going to be junk-mail"
- $P(E)$: "How likely are these words to appear in a randomly selected e-mail"

Bayes classifiers

Principle of Bayes classifiers:

The value of the nominal attribute H should be predicted based on the values of the attributes $A_1, \ldots, A_m$, i.e. the attribute vector $E = (a_1, \ldots, a_m)$.

If h is one of the possible values of attribute H and the other attribute have taken the values $A_1 = a_1, \ldots, A_m = a_m$, then Bayes' theorem yields the probability for $H = h$ given $A_1 = a_1, \ldots, A_m = a_m$:

$$P(H = h | E = (a_1, \ldots, a_m)) = \frac{P(E = (a_1, \ldots, a_m) | H = h) \cdot P(H = h)}{P(E = (a_1, \ldots, a_m))}$$

Bayes classifiers

Compute this probability for all possible values (classes) h of the nominal attribute H and choose the class with the highest probability. (A cost matrix can also be incorporated.)

Since the denominator is independent of h, it does not have any influence on the decision for the class.

Therefore, usually only the likelihoods

$$P(E = (a_1, \dots, a_m) | H = h) \cdot P(H = h)$$

are considered.

Bayes classifiers

The probability $P(H = h)$ can be estimated based on a given data:

$$P(H = h) \; = \; \frac{\text{no. of data from class } h}{\text{no. of data}}$$

In principle, the probability $P(E = (a_1, \ldots, a_m)|H = h)$ could be determined analogously:

$$P(E = (a_1, \ldots, a_m)|H = h) \; =$$
$$\frac{\text{no. of data from class } h \text{ with values } (a_1, \ldots, a_m)}{\text{no. of data from class } h}$$

Bayes classifiers

For $n = 10$ nominal attributes $A_1, \ldots, A_{10}$, each having three possible values, we would need $3^{10} = 59049$ data objects to have at least one example per combination.

Therefore, the computation is carried out under the (naïve, unrealistic) assumption that the attributes $A_1, \ldots, A_m$ are independent given the class, i.e.

$$P(E = (a_1, \ldots, a_m) | H = h) =$$
$$P(A_1 = a_1 | H = h) \cdot \ldots \cdot P(A_m = a_m | H = h)$$

Bayes classifiers

$P(A_i = a_i | H = h)$ can be computed easily:

$$P(A_i = a_i | H = h) = \frac{\text{no. of data from class } h \text{ with } A_i = a_i}{\text{no. of data from class } h}$$

Naïve Bayes classifier

given: A data set with only nominal attributes.

Based on the values $a_1, \ldots, a_m$ of the attributes $A_1, \ldots, A_m$ a prediciton for the value of the attribute H should be derived.

For each class (each value in the domain of H) compute the likelihood

$$L(H = h | A_1 = a_1, \ldots, A_m = a_m) \; =$$

$$P(A_1 = a_1 | H = h) \cdot \ldots \cdot P(A_m = a_m | H = h) \cdot P(H = h)$$

under the assumption that the $A_1, \ldots, A_m$ are independent given the class.

Naïve Bayes classifier

Assign $(A_1, \ldots, a_m)$ to the class h with the highest likelihood.

This Bayes classifier is called naïve because of the (conditional) independence assumption for the attributes $X_1, \ldots, X_m$.

Although this assumption is unrealistic in most cases, the classifier often yields good results, when not too many attributes are correlated.

Example

How does a naïve Bayes classifier classify the object (t, l, y)?

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

We need to calculate

$$L(\text{Sex} = m | \quad \text{Height} = t, \text{Weight} = l, \quad \text{long\_hair} = y)$$

$$= \quad P(\text{Height} = t|\text{Sex} = m) \cdot P(\text{Weight} = l|\text{Sex} = m) \cdot P(\text{long\_hair} = y|\text{Sex} = m) \cdot P(\text{Sex} = m)$$

and

Example

$$
\begin{aligned}
L(\text{Sex} = f| \quad & \text{Height} = t, \\
& \text{Weight} = l,\ \text{Long\_hair} = y) \\
= \quad & P(\text{Height} = t|\text{Sex} = f)\cdot \\
& P(\text{Weight} = l|\text{Sex} = f)\cdot \\
& P(\text{Long\_hair} = y|\text{Sex} = f)\cdot \\
& P(\text{Sex} = f).
\end{aligned}
$$

Example

$P(\text{Height} = t | \text{Sex} = m)$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Height} = t|\text{Sex} = m)$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Height} = t | \text{Sex} = m) = 2/4 = 1/2$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$$P(\text{Weight} = l | \text{Sex} = m) = 0/4 = 0$$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$$P(\text{Long\_hair} = y | \text{Sex} = m) = 0/4 = 0$$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Sex} = m) = 4/10 = 2/5$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$$
\begin{aligned}
L(\mathrm{Sex} = m| \quad & \mathrm{Height} = t, \\
& \mathrm{Weight} = l,\ \mathrm{Long\_hair} = y) \\
= \ & \frac{1}{2} \cdot 0 \cdot 0 \cdot \frac{2}{5} \ = \ 0
\end{aligned}
$$

Example

$P(\text{Height} = t | \text{Sex} = f)$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Height} = t | \text{Sex} = f)$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Height} = t | \text{Sex} = f) = 1/6$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Weight} = l | \text{Sex} = f) = 3/6 = 1/2$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	g	n	n	m

Example

$$P(\text{Long\_hair} = y | \text{Sex} = f) = 4/6 = 2/3$$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$P(\text{Sex} = f) = 6/10 = 3/5$

ID	Height	Weight	Long hair	Sex
1	m	n	n	m
2	s	l	y	f
3	t	h	n	m
4	s	n	y	f
5	t	n	y	f
6	s	l	n	f
7	s	h	n	m
8	m	n	n	f
9	m	l	y	f
10	t	n	n	m

Example

$$\begin{aligned} L(\text{Sex} = f \mid \; & \text{Height} = t, \\ & \text{Weight} = l, \; \text{Long\_hair} = y) \end{aligned}$$

$$= \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{5} = 1/30$$

$$> 0$$

$$\begin{aligned} = L(\text{Sex} = m \mid \; & \text{Height} = t, \\ & \text{Weight} = l, \; \text{Long\_hair} = y) \end{aligned}$$

Classification of (t, l, y): female (f)

Example

The object (t, l, y) was classified by the naïve Bayes classifier.

The data set does not contain any object with this combination of values.

A full Bayes classifier would not be able to classify this object.

More examples

Input	$L(m\vert\ldots)$	$L(f\vert\ldots)$	Class
(t,h,y)	$\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{0}{4}\cdot\frac{4}{10}=0$	$\frac{1}{6}\cdot\frac{0}{6}\cdot\frac{4}{6}\cdot\frac{6}{10}=0$	?
(m,n,n)	$\frac{1}{4}\cdot\frac{2}{4}\cdot\frac{4}{4}\cdot\frac{4}{10}=\frac{1}{20}$	$\frac{2}{6}\cdot\frac{3}{6}\cdot\frac{2}{6}\cdot\frac{6}{10}=\frac{1}{30}$	m
(t,h,n)	$\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{4}{4}\cdot\frac{4}{10}=\frac{1}{10}$	$\frac{1}{6}\cdot\frac{0}{6}\cdot\frac{2}{6}\cdot\frac{6}{10}=0$	m

The object (m,n,n) (Height, Weight, Long_hair) is classified by the naïve Bayes classifier as m although the data sets contains two such objects, one from class m and one from class f.

The main impact comes from the attribut $Long\_hair = n$, having probability 1 in class m, but a low probability in class f.

Laplace correction

If a single likelihood is zero, then the overall likelihood is zero automatically, even the when the other likelihoods are high.

Therefore: Laplace correction:

$$\hat{P}(A_k = a_k \mid C = c_i) = \frac{\#(A_k = a_k, C = c_i) + \gamma}{\#(C = c_i) + n_{A_k}\gamma}$$

γ is called Laplace correction.

$\gamma = 0$: Maximum likelihood estimation.

Common choices: $\gamma = 1$ or $\gamma = \frac{1}{2}$.

Laplace correction

Laplace correction for $P(\text{Height} = \ldots|\text{Sex} = m)$ with $\gamma = 1$

Height	#	Laplace	P	P_{Laplace}
s	1	2	1/4	2/7
m	1	2	1/4	2/7
t	2	3	2/4	3/7

Naïve Bayes classifier: Implementation

The counting of the frequencies should be carried out once when the naïve Bayes classifier is constructed.

The probability distribution for the single attributes should be stored in a table.

When the naïve Bayes classifier is applied to new data, only the corresponding values in the table are needed.

Time complexity of training naïve Bayes: $O(nk)$ where n number of training instance and k average number of non-zero features for an instance (scales to large data sets). Complexity of applying the model on new instance: $O(k)$.

Treatment of missing values

During learning: The missing values are simply not counted for the frequencies of the corresponding attribute.

During classification: Only the probabilities (likelihoods) of those attributes are multiplied for which a value is available.

Naïve Bayes classifier: Implementation

The likelihood for class C_i is defined as

$$P(C = C_i) \cdot \prod_{j=1}^{d} P(A_j = a_j | C = C_i)$$

and we predict the class with highest likelihood. However, if d is large, we will have numerical problems in computations (floating point arithmetics underflow). Therefore, we rather compare the logarithms of the likelihoods for each class:

$$log(P(C = C_i)) + \sum_{j=i}^{d} log(P(A_j = a_j | C = C_i))$$

(Remember, $log(a \cdot b) = log(a) + log(b)$)

Numerical attributes

Estimation of probabilities:

- **Numerical attributes:** Assume a normal distribution.

$$f(X_k = x_k \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_k(c_i)} \exp\left(-\frac{(x_k - \mu_k(c_i))^2}{2\sigma_k^2(c_i)}\right)$$

- Estimation of the mean value

$$\hat{\mu}_k(c_i) = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C=c_i)} x_k(j)$$

Numerical attributes

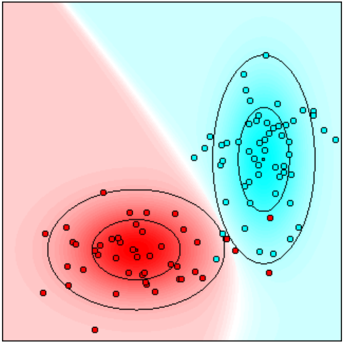
- Estimation of the variance

$$\hat{\sigma}_k^2(c_i) = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} (x_k(j) - \hat{\mu}_k(c_i))^2$$

$\xi = \#(C = c_i)$: Maximum likelihood estimation
$\xi = \#(C = c_i) - 1$: Unbiased estimation

Example

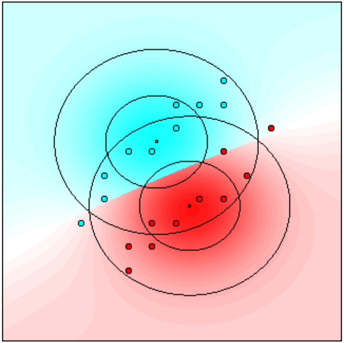
- 100 data points, 2 classes
- Small squares: mean values
- Inner ellipses:
  one standard deviation
- Outer ellipses:
  two standard deviations
- Classes overlap:
  classification is not perfect



Naïve Bayes classifier

Example

- 20 data points, 2 classes
- Small squares: mean values
- Inner ellipses:
  one standard deviation
- Outer ellipses:
  two standard deviations
- Attributes are not conditionally independent given the class



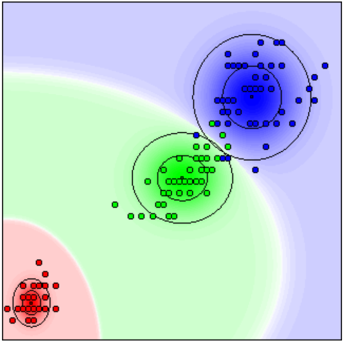
Naïve Bayes classifier

Example: Iris data

- 150 data points, 3 classes

  Iris setosa (red)
  Iris versicolor (green)
  Iris virginica (blue)
- Shown: 2 out of 4 attributes

  sepal length
  sepal width
  petal length (horizontal)
  petal width (vertical)
- 6 misclassifications on the training data (with all 4 attributes)



Naïve Bayes classifier

Full Bayes classifiers

- Often restricted to metric/numeric attributes (only the class is nominal/symbolic).
- **Simplifying Assumption:**
  Each class can be described by a multivariate normal distribution.

$$
\begin{aligned}
& f(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \\
& = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2} (\mathbf{a} - \mu_{\mathbf{i}})^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{a} - \mu_i) \right)
\end{aligned}
$$

μ_i: mean value vector for class c_i
$\boldsymbol{\Sigma}_i$: covariance matrix for class c_i

Full Bayes classifiers

- Intuitively: Each class has a bell-shaped probability density.
- Naive Bayes classifiers: Covariance matrices are diagonal matrices.
  (Details about this relation are given below.)

Full Bayes classifiers

Estimation of Probabilities:

- Estimation of the mean value vector

$$\hat{\mu}_i = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C=c_i)} \mathbf{a}(j)$$

- Estimation of the covariance matrix

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} (\mathbf{a}(j) - \hat{\mu}_i)(\mathbf{a}(j) - \hat{\mu}_i)^\top$$

$\xi = \#(C = c_i)$: Maximum likelihood estimation
$\xi = \#(C = c_i) - 1$: Unbiased estimation

Naïve vs. full Bayes classifiers

Assuming that covariance matrices are diagonal matrices:

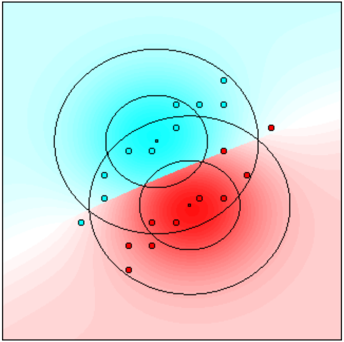
$$
\begin{aligned}
& f(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \\
&= \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}_i|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{a} - \mu_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{a} - \mu_i)\right) \\
&= \frac{1}{\sqrt{(2\pi)^m \prod_{k=1}^m \sigma_{i,k}^2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{a} - \mu_i)^\top \operatorname{diag}\left(\sigma_{i,1}^{-2}, \ldots, \sigma_{i,m}^{-2}\right)(\mathbf{a} - \mu_i)\right) \\
&= \frac{1}{\prod_{k=1}^m \sqrt{2\pi\sigma_{i,k}^2}} \cdot \exp\left(-\frac{1}{2}\sum_{k=1}^m \frac{(a_k - \mu_{i,k})^2}{\sigma_{i,k}^2}\right) \\
&= \prod_{k=1}^m \frac{1}{\sqrt{2\pi\sigma_{i,k}^2}} \cdot \exp\left(-\frac{(a_k - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right) \;\widehat{=}\; \prod_{k=1}^m f(A_k = a_k \mid C = c_i),
\end{aligned}
$$

where $f(A_k = a_k \mid C = c_i)$ are the density functions used by a naïve Bayes classifier.

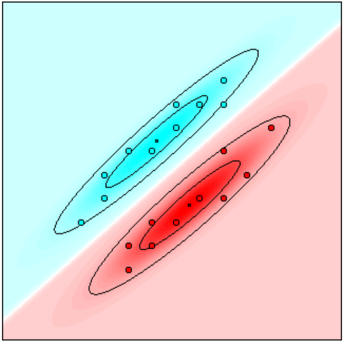
Naïve vs. full Bayes classifiers

Naïve Bayes classifiers for numerical data are equivalent to full Bayes classifiers with diagonal covariance matrices.

Naïve vs. full Bayes classifiers



Naïve Bayes classifier



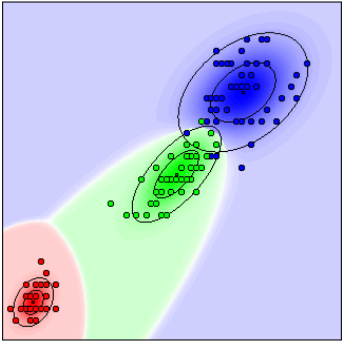
Full Bayes classifier

Full Bayes classifier: Iris data

- 150 data points, 3 classes

  Iris setosa (red)
  Iris versicolor (green)
  Iris virginica (blue)
- Shown: 2 out of 4 attributes

  sepal length
  sepal width
  petal length (horizontal)
  petal width (vertical)
- 2 misclassifications on the training data (with all 4 attributes)



Full Bayes classifier