Statistics and learning
Naive Bayes Classifiers

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Friday 22\textsuperscript{nd} November 2013
A bit of intuition

<table>
<thead>
<tr>
<th>S(ex)</th>
<th>H(eight) (m)</th>
<th>W(eight) (kg)</th>
<th>F(oot size) (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>1.82</td>
<td>82</td>
<td>30</td>
</tr>
<tr>
<td>M</td>
<td>1.80</td>
<td>86</td>
<td>28</td>
</tr>
<tr>
<td>M</td>
<td>1.70</td>
<td>77</td>
<td>30</td>
</tr>
<tr>
<td>M</td>
<td>1.80</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>F</td>
<td>1.52</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>F</td>
<td>1.65</td>
<td>68</td>
<td>20</td>
</tr>
<tr>
<td>F</td>
<td>1.68</td>
<td>59</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>1.75</td>
<td>68</td>
<td>23</td>
</tr>
</tbody>
</table>

Is (1.81, 59, 21) male or female?
Bayesian probabilities

Question: $P(S = M| (H, W, F) = (1.81, 59, 21)) > P(S = F| (H, W, F) = (1.81, 59, 21))$?
Bayesian probabilities

**Question:** \( \mathbb{P}(S = M \mid (H, W, F) = (1.81, 59, 21)) > \mathbb{P}(S = F \mid (H, W, F) = (1.81, 59, 21)) \)?

**Bayes law:**

\[
\mathbb{P}(S \mid H, W, F) = \frac{\mathbb{P}(S) \times \mathbb{P}(H, W, F \mid S)}{\mathbb{P}(H, W, F)}
\]

In other words:

posterior = \( \text{prior} \times \text{likelihood} \) \( \text{evidence} \)

**But** \( \mathbb{P}(H, W, F) \) does not depend on \( S \), so the question boils down to:

\[
\mathbb{P}(S = M) \times \mathbb{P}(H, W, F \mid S = M) > \mathbb{P}(S = F) \times \mathbb{P}(H, W, F \mid S = F)
\]

\( \mathbb{P}(S) \) is easy to estimate. What about \( \mathbb{P}(H, W, F \mid S) \)?
Bayesian probabilities

Question: $P(S = M \mid (H, W, F)) = (1.81, 59, 21)) > P(S = F \mid (H, W, F) = (1.81, 59, 21))$?

Bayes law:

$$P(S \mid H, W, F) = \frac{P(S) \times P(H, W, F \mid S)}{P(H, W, F)}$$

In other words:

posterior = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}

But $P(H, W, F)$ does not depend on $S$, so the question boils down to:

$$P(S = M) \times P(H, W, F \mid S = M) > P(S = F) \times P(H, W, F \mid S = F)?$$

$P(S)$ is easy to estimate. What about $P(H, W, F \mid S)$?
Naive Bayes hypothesis

Discretize $range(H)$ in 10 segments.

$\rightarrow \mathbb{P}(H, W, F|S)$ is a 3-dimensional array.

$\rightarrow 10^3$ values to estimate

$\rightarrow$ requires lots of data!
Naive Bayes hypothesis

Discretize $\text{range}(H)$ in 10 segments.

$\rightarrow \mathbb{P}(H, W, F|S)$ is a 3-dimensional array.

$\rightarrow 10^3$ values to estimate

$\rightarrow$ requires lots of data!

**Curse of dimensionality:** $(\text{data})$ scales exponentially with $(\text{features})$. 

$E. \text{Rachelson \ & M. Vignes (ISAE)}$

SAD 2013 4 / 16
**Naive Bayes hypothesis**

Discretize \( \text{range}(H) \) in 10 segments.
- \( \mathbb{P}(H, W, F|S) \) is a 3-dimensional array.
- \( 10^3 \) values to estimate
- requires lots of data!

**Curse of dimensionality:** \( \#(\text{data}) \) scales exponentially with \( \#(\text{features}) \).

Reminder, conditional probabilities:

\[
\mathbb{P}(H, W, F|S) = \mathbb{P}(H|S) \times \mathbb{P}(W|S,H) \times \mathbb{P}(F|S,H,W)
\]
Naive Bayes hypothesis

Discretize \( \text{range}(H) \) in 10 segments.
\[ \rightarrow P(H, W, F|S) \] is a 3-dimensional array.
\[ \rightarrow 10^3 \text{ values to estimate} \]
\[ \rightarrow \text{requires lots of data!} \]

**Curse of dimensionality:** \( \#(\text{data}) \) scales exponentially with \( \#(\text{features}) \).

Reminder, conditional probabilities:

\[ P(H, W, F|S) = P(H|S) \times P(W|S, H) \times P(F|S, H, W) \]

Naive Bayes: “what if
\[ \begin{align*}
P(W|S, H) &= P(W|S) \\
P(F|S, H, W) &= P(F|S)
\end{align*} \]?

\[ \rightarrow \text{Then } P(H, W, F|S) = P(H|S) \times P(W|S) \times P(F|S) \]
\[ \rightarrow \text{only } 3 \times 10 \text{ values to estimate} \]
Naive Bayes hypothesis, cont’d

\[ P(W \mid S, H) = P(W \mid S) \]

what does that mean?

“Among male individuals, the weight is independent of the height”

What do you think?
Naive Bayes hypothesis, cont’d

\[ P(W|S, H) = P(W|S) \]  what does that mean?

“Among male individuals, the weight is \textit{independent} of the height”

What do you think?

Despite that \textit{naive} assumption, Naive Bayes classifiers perform very well!
Naive Bayes hypothesis, cont’d

\[ \mathbb{P}(W|S, H) = \mathbb{P}(W|S) \]

what does that mean?

“Among male individuals, the weight is \textit{independent} of the height”

What do you think?

Despite that \textit{naive} assumption, Naive Bayes classifiers perform very well!

Let’s formalize that a little more.
Naive Bayes classifiers in one slide!

\[ \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \]
Naive Bayes classifiers in one slide!

\[
\text{posterior} = \frac{\text{prior } \times \text{ likelihood}}{\text{evidence}}
\]

\[
\mathbb{P}(Y|X_1, \ldots, X_n) = \frac{\mathbb{P}(Y) \times \mathbb{P}(X_1, \ldots, X_n|Y)}{\mathbb{P}(X_1, \ldots, X_n)}
\]
Naive Bayes classifiers in one slide!

\[ \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \]

\[ P(Y|X_1, \ldots, X_n) = \frac{P(Y) \times P(X_1, \ldots, X_n|Y)}{P(X_1, \ldots, X_n)} \]

Naive conditional independence assump.: \( \forall i \neq j, P(X_i|Y, X_j) = P(X_i|Y) \)
Naive Bayes classifiers in one slide!

\[ \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \]

\[ \mathbb{P}(Y|X_1, \ldots, X_n) = \frac{\mathbb{P}(Y) \times \mathbb{P}(X_1, \ldots, X_n|Y)}{\mathbb{P}(X_1, \ldots, X_n)} \]

Naive conditional independence assump.: \( \forall i \neq j, \mathbb{P}(X_i|Y, X_j) = \mathbb{P}(X_i|Y) \)

\[ \Rightarrow \mathbb{P}(Y|X_1, \ldots, X_n) = \frac{1}{\mathbb{Z}} \times \mathbb{P}(Y) \times \prod_{i=1}^{n} \mathbb{P}(X_i|Y) \]
Naive Bayes classifiers in one slide!

\[ \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \]

\[ \mathbb{P}(Y|X_1,\ldots,X_n) = \frac{\mathbb{P}(Y) \times \mathbb{P}(X_1,\ldots,X_n|Y)}{\mathbb{P}(X_1,\ldots,X_n)} \]

Naive conditional independence assump.: \( \forall i \neq j, \mathbb{P}(X_i|Y, X_j) = \mathbb{P}(X_i|Y) \)

\[ \Rightarrow \mathbb{P}(Y|X_1,\ldots,X_n) = \frac{1}{Z} \times \mathbb{P}(Y) \times \prod_{i=1}^{n} \mathbb{P}(X_i|Y) \]

If \( \left\{ Y \in \{1,\ldots,k\} \right\} \), the NBC has \((k-1) + nqk\) parameters \(\theta\).

Given \( \{x_i, y_i\}_{0 \leq i \leq N} \), \(\theta = \hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} (\log) L(x_1 \ldots x_N; \theta)\)

prediction: NBC \( (x) := \arg\max_{y \in \{1,\ldots,k\}} \mathbb{P}(Y = y|\hat{\theta}) \times \prod_{i=1}^{n} \mathbb{P}(X_i|Y = y) \)
Naive Bayes classifiers in one slide!

\[ \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \]

\[
\mathbb{P}(Y|X_1, \ldots, X_n) = \frac{\mathbb{P}(Y) \times \mathbb{P}(X_1, \ldots, X_n|Y)}{\mathbb{P}(X_1, \ldots, X_n)}
\]

Naive conditional independence assump.: \( \forall i \neq j, \mathbb{P}(X_i|Y, X_j) = \mathbb{P}(X_i|Y) \)

\[
\Rightarrow \mathbb{P}(Y|X_1, \ldots, X_n) = \frac{1}{\mathcal{Z}} \times \mathbb{P}(Y) \times \prod_{i=1}^{n} \mathbb{P}(X_i|Y)
\]

If \( \left\{ Y \in \{1, \ldots, k\} \right\} \) \( \mathbb{P}(X_i|Y) \sim q \) params

Given \( \{x_i, y_i\}_{0 \leq i \leq N} \), \( \theta = \hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} (\log)L(x_1 \ldots x_N; \theta) \)

Prediction: \( \text{NBC}(x) := \arg\max_{y \in [1,k]} \mathbb{P}_{\hat{\theta}}(Y = y) \times \prod_{i=1}^{n} \mathbb{P}_{\hat{\theta}}(X_i = x_i|Y = y) \)
\[ \mathbb{P}(S|H, W, F) = \frac{1}{Z} \times \mathbb{P}(S) \times \mathbb{P}(H|S) \times \mathbb{P}(W|S) \times \mathbb{P}(F|S) \]

<table>
<thead>
<tr>
<th>S(ex)</th>
<th>H(eight) (m)</th>
<th>W(eight) (kg)</th>
<th>F(oot size) (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>1.82</td>
<td>82</td>
<td>30</td>
</tr>
<tr>
<td>M</td>
<td>1.80</td>
<td>86</td>
<td>28</td>
</tr>
<tr>
<td>M</td>
<td>1.70</td>
<td>77</td>
<td>30</td>
</tr>
<tr>
<td>M</td>
<td>1.80</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>F</td>
<td>1.52</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>F</td>
<td>1.65</td>
<td>68</td>
<td>20</td>
</tr>
<tr>
<td>F</td>
<td>1.68</td>
<td>59</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>1.75</td>
<td>68</td>
<td>23</td>
</tr>
</tbody>
</table>

\[ \mathbb{P}(S = M) = ? \]
\[ \mathbb{P}(H = 1.81|S = M) = ? \]
\[ \mathbb{P}(W = 59|S = M) = ? \]
\[ \mathbb{P}(F = 21|S = M) = ? \]
Back to the example

\[
P(S|H,W,F) = \frac{1}{Z} \times P(S) \times P(H|S) \times P(W|S) \times P(F|S)
\]

```r
> gens <- read.table("sex_classif.csv", sep=";", colnames)
> library("MASS")
> fitdistr(gens[1:4,2],"normal")
...
> 0.5*dnorm(1.81,mean=1.78,sd=0.04690416)
*dnorm(59,mean=80,sd=4.301163)
*dnorm(21,mean=28.25,sd=2.0463382)
> 0.5*dnorm(1.81,mean=1.65,sd=0.08336666)
*dnorm(59,mean=60,sd=9.407444)
*dnorm(21,mean=19,sd=2.915476)
```
Back to the example

\[ \mathbb{P}(S|H, W, F) = \frac{1}{Z} \times \mathbb{P}(S) \times \mathbb{P}(H|S) \times \mathbb{P}(W|S) \times \mathbb{P}(F|S) \]

\( S \) is discrete, \( H, W \) and \( F \) are assumed Gaussian.

| \( S \) | \( \hat{p}_S \) | \( \hat{\mu}_{H|S} \) | \( \hat{\sigma}_{H|S} \) | \( \hat{\mu}_{W|S} \) | \( \hat{\sigma}_{W|S} \) | \( \hat{\mu}_{F|S} \) | \( \hat{\sigma}_{F|S} \) |
|-------|------|------|------|------|------|------|------|
| \( M \) | 0.5  | 1.78 | 0.0469 | 80   | 4.3012 | 28.25 | 2.0463 |
| \( F \) | 0.5  | 1.65 | 0.0834 | 60   | 9.4074 | 19    | 2.9154 |

\[ \mathbb{P}(M|1.81, 59, 21) = \frac{1}{Z} \times 0.5 \times \frac{e^{-(1.78-1.81)^2 / 2 \cdot 0.0469^2}}{\sqrt{2\pi} 0.0469^2} \times \frac{\sqrt{2\pi} 0.0469^2}{\sqrt{2\pi} 4.3012^2} \times \frac{\sqrt{2\pi} 4.3012^2}{\sqrt{2\pi} 2.0463^2} \]

\[ \mathbb{P}(F|1.81, 59, 21) = \frac{1}{Z} \times 0.5 \times \frac{e^{-(1.65-1.81)^2 / 2 \cdot 0.0834^2}}{\sqrt{2\pi} 0.0834^2} \times \frac{\sqrt{2\pi} 0.0834^2}{\sqrt{2\pi} 9.4074^2} \times \frac{\sqrt{2\pi} 9.4074^2}{\sqrt{2\pi} 2.9154^2} \]

\[ \mathbb{P}(M|1.81, 59, 21) = \frac{1}{Z} \times 7.854 \cdot 10^{-10} \]

\[ \mathbb{P}(F|1.81, 59, 21) = \frac{1}{Z} \times 1.730 \cdot 10^{-3} \]
Back to the example

\[
\mathbb{P}(S|H, W, F) = \frac{1}{Z} \times \mathbb{P}(S) \times \mathbb{P}(H|S) \times \mathbb{P}(W|S) \times \mathbb{P}(F|S)
\]

Conclusion: given the data, (1.81m, 59kg, 21cm) is more likely to be female.
Features

\[ \mathbb{P}(Y|X_1, \ldots, X_n) = \frac{1}{Z} \times \mathbb{P}(Y) \times \prod_{i=1}^{n} \mathbb{P}(X_i|Y) \]

Continuous \( X_i \)

- Assume normal distribution \( X_i|Y = y \sim \mathcal{N}(\mu_{iy}, \sigma_{iy}) \)
- Discretize \( X_i|Y = y \) via binning (often better if many data points)

Binary \( X_i \)

- Bernouilli distribution \( X_i|Y = y \sim \mathcal{B}(p_{iy}) \)
Algorithm

Train:
For all possible values of $Y$ and $X_i$,
compute $\hat{P}(Y = y)$ and $\hat{P}(X_i = x_i | Y = y)$.

Predict:
Given $(x_1, \ldots, x_n)$, return $y$ that
maximizes $\hat{P}(Y = y) \prod_{i=1}^{n} P(X_i = x_i | Y = y)$. 
When should you use NBC?

- Needs little data to estimate parameters.
- Can easily deal with large feature spaces.
- Requires little tuning (but a bit of feature engineering).
- Without good tuning, more complex approaches are often outperformed by NBC.

...despite the independence assumption!

If you want to understand why:

A little more

Never say never!

\[ \hat{P}(Y = y|X_i = x_i, i \in [1, n]) = \frac{1}{Z} \hat{P}(Y = y) \times \prod_{i=1}^{n} \hat{P}(X_i = x_i|Y = y) \]

But if \( \hat{P}(X_i = x_i|Y = y) = 0 \), then all other info from \( X_j \) is lost!

\[ \rightarrow \text{never set a probability estimate below } \epsilon \text{ (sample correction)} \]

Additive model

Log-likelihood: \( \log \frac{\hat{P}(Y|X)}{\hat{P}(\bar{Y}|X)} = \log \frac{\hat{P}(Y)}{1 - \hat{P}(Y)} + \sum_{i=1}^{n} \log \frac{\hat{P}(X_i|Y)}{\hat{P}(X_i|\bar{Y})} \)

\[ = \alpha + \sum_{i=1}^{n} g(X_i) \]
A real-world example: spam filtering

Build a NBC that classifies emails as spam/non-spam, using the occurrence of words.

Any ideas?
The data

Data = a bunch of emails, labels as spam/non-spam.
The Ling-spam dataset:

Preprocessing
Form each email, remove:

- stop-words
- lemmatization
- non-words
Before:
Subject: Re: 5.1344 Native speaker intuitions
The discussion on native speaker intuitions has been extremely interesting, but I worry that my brief intervention may have muddied the waters. I take it that there are a number of separable issues. The first is the extent to which a native speaker is likely to judge a lexical string as grammatical or ungrammatical per se. The second is concerned with the relationships between syntax and interpretation (although even here the distinction may not be entirely clear cut).

After:
re native speaker intuition discussion native speaker intuition extremely interest worry brief intervention muddy waters number separable issue first extent native speaker likely judge lexical string grammatical ungrammatical per se second concern relationship between syntax interpretation although even here distinction entirely clear cut
The data

- Keep a dictionary $V$ of the $|V|$ most frequent words.
- Count the occurrence of each dictionary word in each example email.

- $m$ emails
- $n_i$ words in email $i$
- $|V|$ words in dictionary

- What is $Y$?
- What are the $X_i$?
Text classification features

\[ Y = 1 \text{ if the email is a spam.} \]

\[ X_k = 1 \text{ if word } i \text{ of dictionary appears in the email} \]

Estimator of \( P(X_k = 1|Y = y) \):
\[ x_{ij}^i \text{ is the } j \text{th word of email } i, \ y^i \text{ is the label of email } i. \]

\[ \phi_{ky} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_{ij}^i=k \text{ and } y^i=y\} + 1}{\sum_{i=1}^{m} 1\{y^i=y\} n_i + |V|} \]
Getting started in R

```r
> trainingSet <- read.table("emails-train-features.txt", sep=" ",
col.names=c("document","word","count"))
> labelSet <- read.table("emails-train-labels.txt", sep=" ",
col.names=c("spam"))
> num.features <- 2500
> doc.word.train <- spMatrix(max(trainingSet[,1]), num.features,
  as.vector(trainingSet[,1]), as.vector(trainingSet[,2]),
  as.vector(trainingSet[,3]))
> doc.class.train <- labelSet[,1]
> source("trainSpamClassifier") # your very own classifier!
> params <- trainSpamClassifier(doc.word.train,doc.class.train)
> testingSet <- read.table("emails-test-features.txt", sep=" ",
col.names=c("document","word","count"))
> doc.word.test <- spMatrix(max(testingSet[,1]), num.features,
  as.vector(testingSet[,1]), as.vector(testingSet[,2]),
  as.vector(testingSet[,3]))
> source("testSpamClassifier.r")
> prediction <- testSpamClassifier(params, doc.word.test) # does it work well?
```
Going further in text mining in R

The “Text Mining” package:
http://cran.r-project.org/web/packages/tm/
http://tm.r-forge.r-project.org/

Useful if you want to change the features on the previous dataset.