Statistics and learning

Naive Bayes Classifiers

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Friday 22nd November 2013

A bit of intuition

S(ex)	H(eight) (m)	W(eight) (kg)	F(oot size) (cm)
M	1.82	82	30
M	1.80	86	28
M	1.70	77	30
M	1.80	75	25
F	1.52	45	15
F	1.65	68	20
F	1.68	59	18
F	1.75	68	23

Is $(1.81, 59, 21)$ male of female?

Bayesian probabilities

Question: $\mathbb{P}\left(S = M \middle| (H, W, F) = (1.81, 59, 21)\right) >$
$\mathbb{P}\left(S = F \middle| (H, W, F) = (1.81, 59, 21)\right)$?

Bayesian probabilities

Question: $\mathbb{P}\left(S = M \middle| (H, W, F) = (1.81, 59, 21)\right) >$
$\mathbb{P}\left(S = F \middle| (H, W, F) = (1.81, 59, 21)\right)$?

Bayes law:

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

In other words:

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

Bayesian probabilities

Question: $\mathbb{P}\left(S = M \middle| (H, W, F) = (1.81, 59, 21)\right) > \mathbb{P}\left(S = F \middle| (H, W, F) = (1.81, 59, 21)\right)$?

Bayes law:

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

In other words:

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

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

$$\mathbb{P}(S = M) \times \mathbb{P}(H, W, F|S = M) > \mathbb{P}(S = F) \times \mathbb{P}(H, W, F|S = F)?$$

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

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 $range(H)$ 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: #(data) scales exponentially with #(features).

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!

Curse of dimensionality: #(data) scales exponentially with #(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 $range(H)$ 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: #(data) scales exponentially with #(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: "what if $\left\{ \begin{array}{rcl} \mathbb{P}(W|S, H) & = & \mathbb{P}(W|S) \\ \mathbb{P}(F|S, H, W) & = & \mathbb{P}(F|S) \end{array} \right.$?"

$\rightarrow$ Then $\mathbb{P}(H, W, F|S) = \mathbb{P}(H|S) \times \mathbb{P}(W|S) \times \mathbb{P}(F|S)$

$\rightarrow$ only 3×10 values to estimate

Naive Bayes hypothesis, cont'd

$\mathbb{P}(W|S,H) = \mathbb{P}(W|S)$ what does that mean?
"Among male individuals, the weight is *independent* of the height"
What do you think?

Naive Bayes hypothesis, cont'd

$\mathbb{P}(W|S,H) = \mathbb{P}(W|S)$ what does that mean?
"Among male individuals, the weight is *independent* of the height"
What do you think?

Despite that *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 *independent* of the height"
What do you think?

Despite that *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!

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

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

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

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.: $\forall i \neq j, \mathbb{P}(X_i|Y, X_j) = \mathbb{P}(X_i|Y)$

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

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 $\begin{cases} Y \in \{1, \ldots, k\} \\ \mathbb{P}(X_i|Y) \sim q \text{ params} \end{cases}$, the NBC has $(k-1) + nqk$ parameters θ.

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

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 $\begin{cases} Y \in \{1, \ldots, k\} \\ \mathbb{P}(X_i|Y) \sim q \text{ params} \end{cases}$, the NBC has $(k-1) + nqk$ parameters θ.

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

Prediction: $\text{NBC}(x) := \underset{y \in [1,k]}{\text{argmax}} \, \mathbb{P}_{\hat{\theta}}(Y = y) \times \prod_{i=1}^{n} \mathbb{P}_{\hat{\theta}}(X_i = x_i | Y = y)$

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(ex)	H(eight) (m)	W(eight) (kg)	F(oot size) (cm)
M	1.82	82	30
M	1.80	86	28
M	1.70	77	30
M	1.80	75	25
F	1.52	45	15
F	1.65	68	20
F	1.68	59	18
F	1.75	68	23

$$\mathbb{P}(S = M) = ?$$
$$\mathbb{P}(H = 1.81|S = M) = ?$$
$$\mathbb{P}(W = 59|S = M) = ?$$
$$\mathbb{P}(F = 21|S = M) = ?$$

Back to the example

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

```
> 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

$$\begin{aligned}\mathbb{P}(M|1.81,59,21) &= \frac{1}{Z} \times 0.5 \times \frac{e^{\frac{-(1.78-1.81)^2}{2\cdot 0.0469^2}}}{\sqrt{2\pi 0.0469^2}} \times \frac{e^{\frac{-(80-59)^2}{2\cdot 4.3012^2}}}{\sqrt{2\pi 4.3012^2}} \times \frac{e^{\frac{-(28.25-21)^2}{2\cdot 2.0463^2}}}{\sqrt{2\pi 2.0463^2}} \\ &= \frac{1}{Z} \times 7.854 \cdot 10^{-10}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(F|1.81,59,21) &= \frac{1}{Z} \times 0.5 \times \frac{e^{\frac{-(1.65-1.81)^2}{2\cdot 0.0834^2}}}{\sqrt{2\pi 0.0834^2}} \times \frac{e^{\frac{-(60-59)^2}{2\cdot 9.4074^2}}}{\sqrt{2\pi 9.4074^2}} \times \frac{e^{\frac{-(19-21)^2}{2\cdot 2.9154^2}}}{\sqrt{2\pi 2.9154^2}} \\ &= \frac{1}{Z} \times 1.730 \cdot 10^{-3}\end{aligned}$$

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

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

<u>Predict</u>:
Given $(x_1, \ldots, x_n)$, return y that
maximizes $\hat{\mathbb{P}}(Y = y) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i | Y = y)$.

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:

The Optimality of Naive Bayes, H. Zhang, *FLAIRS*, 2004.

A little more

Never say never!

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

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

$\rightarrow$ never set a probability estimate below ϵ (sample correction)

Additive model

Log-likelihood: $\log \hat{\mathbb{P}}(Y|X) = -\log Z + \log \hat{\mathbb{P}}(Y) + \sum_{i=1}^{n} \log \hat{\mathbb{P}}(X_i|Y)$ and:

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

$$= \alpha + \sum_{i=1}^{n} g(X_i)$$

A real-world example: spam filtering

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

Any ideas?

The data

Data = a bunch of emails, labels as spam/non-spam.
The Ling-spam dataset:
`http://csmining.org/index.php/ling-spam-datasets.html`.

Preprocessing

Form each email, remove:

- stop-words
- lemmatization
- non-words

The data

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 dictionnary V of the $|V|$ most frequent words.
- Count the occurence 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 $\mathbb{P}(X_k = 1 | Y = y)$:
x_j^i is the jth word of email i, y^i is the label of email i.

$$\phi_{ky} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1_{\{x_j^i = k \text{ and } y^i = y\}} + 1}{\sum_{i=1}^{m} 1_{\{y^i = y\}} n_i + |V|}$$

Getting started in 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.