Naïve Bayesian Classifier (NBC)

Lesson 3 – Section 4

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Recap

>Linear regression (review)

>Logistic regression (review)

>K-nearest neighbors, pros and cons

Overview

Mathematical formula of Naïve Bayesian Classifier

How to derive NBC from data

Algorithm of NBC for discrete variables

Algorithm of NBC for continuous variables

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Naïve Bayesian Classifier

Naïve Bayesian Classifier

$$Pr(yi = c|\boldsymbol{X}_i) = \frac{\Pr(yi = c, \boldsymbol{X}_i)}{\Pr(\boldsymbol{X}_i)}$$
$$= \frac{\Pr(\boldsymbol{X}_i|yi{=}c)\cdot\Pr(y_i{=}c)}{\Pr(\boldsymbol{X}_i)}$$

- $\Pr(\boldsymbol{X}_i)$ is a common factor for all classes
- Only need to compare $\Pr(\boldsymbol{X}_i|y_i = c) \cdot \Pr(y_i = c)$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

How to Derive Likelihood from Data?

$$Pr(\boldsymbol{X}_i|y_i = c) = \Pr(x_{i1}, x_{i2}, \dots, x_{id}|\ y_i = c)$$

Naïve Bayesian Classification: Assuming independence among $x_{i1}, x_{i2}, \dots, x_{id}$ condition on $y_i = c$

$$\begin{aligned}\Pr(x_{i1}, x_{i2}, \dots, x_{id}|\ y_i = c) &= \Pr(x_{i1}|\ x_{i2}, \dots, x_{id}, y_i = c)\Pr(x_{i2}, \dots, x_{id}|\ y_i = c)\\ &= \Pr(x_{i1}|\ y_i = c)\ \Pr(x_{i2}, \dots, x_{id}|\ y_i = c)\\ &= \dots\\ &= \Pr(x_{i1}|\ y_i = c)\ \Pr(x_{i2}|\ y_i = c) \dots \Pr(x_{id}|\ y_i = c)\end{aligned}$$

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NBC for Discrete Features

- Algorithm: Discrete-Valued Features
  - Learning Phase: Given a training set **S** of F features and L classes,

For each target value of c_i $(c_i = c_1, \cdots, c_L)$

$\hat{P}(c_i) \leftarrow$ estimate $P(c_i)$ with examples in S;

For every feature value x_{jk} of each feature x_j $(j = 1, \cdots, F;\ k = 1, \cdots, N_j)$

$\hat{P}(x_j = x_{jk} \mid c_i) \leftarrow$ estimate $P(x_{jk} \mid c_i)$ with examples in S;

Output: $F * L$ conditional probabilistic (generative) models

  - Test Phase: Given an unknown instance $\mathbf{x}' = (a'_1, \cdots, a'_n)$

"Look up tables" to assign the label c^* to $\mathbf{X}'$ if

$$[\hat{P}(a'_1 \mid c^*) \cdots \hat{P}(a'_n \mid c^*)]\hat{P}(c^*) > [\hat{P}(a'_1 \mid c_i) \cdots \hat{P}(a'_n \mid c_i)]\hat{P}(c_i), \quad c_i \neq c^*, c_i = c_1, \cdots, c_L$$

Example of NBC

- Tennis.csv

outlook	temp	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Derive Conditional Probabilities, and Prior Probabilities of Each Features: Usage of Training Samples

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=N*o*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

P(Play=*Yes*) = 9/14 *P*(Play=*No*) = 5/14

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Use the Learned Probabilities to Predict Testing Cases

- Consider a testing case:

X=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

$\Pr(y = play|Sunny, Cool, High, Strong) \propto$
Pr(Sunny|play)*Pr(Cool|play)*Pr(High|play)
*Pr(Strong|play)*Pr(play)
=2/9*3/9*3/9*3/9*9/14=0.00529

$\Pr(y = no\ play|Sunny, Cool, High, Strong) \propto$ Pr(Sunny|no play)*Pr(Cool|no play)*Pr(High|no play)
*Pr(Strong|no play)*Pr(no play)
=3/5*1/5*4/5*3/5*5/14=0.02057

So, we should assign label ***No Play*** to this condition.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Algorithm of NBC with Continuous Features

Algorithm: Continuous-valued Features

- Numberless values taken by a continuous-valued feature
- Conditional probability often modeled with the normal distribution

$$\hat{P}(x_j \mid c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (avearage) of feature values x_j of examples for which $c = c_i$

σ_{ji} : standard deviation of feature values x_j of examples for which $c = c_i$

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Algorithm of NBC with Continuous Features

Learning Phase: for $\mathbf{X} = (X_1, \cdots, X_F)$, $C = c_1, \cdots, c_L$

Output: $F \times L$ normal distributions and $P(C = c_i)$ $i = 1, \cdots, L$

Test Phase: Given an unknown instance $\mathbf{X}' = (a_1', \cdots, a_n')$

- Instead of looking-up tables, calculate conditional probabilities with all the normal distributions achieved in the learning phrase
- Apply the MAP rule to assign a label (the same as done for the discrete case)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NBC with Continuous Features Example

Example: Continuous-valued Features

– Temperature is naturally of continuous value.

Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

No: 27.3, 30.1, 17.4, 29.5, 15.1

– Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

NBC with Continuous Features Example

Learning Phase: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2\times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2\times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Zero conditional probability

- If no example contains the feature value
  - In this circumstance, we face a zero conditional probability problem during test

$$\hat{P}(x_1|c_i)\cdots\hat{P}(a_{jk}|c_i)\cdots\hat{P}(x_n|c_i)=0 \qquad \text{for } x_j=a_{jk},\ \hat{P}(a_{jk}|c_i)=0$$

  - For a remedy, class conditional probabilities re-estimated with

$$\hat{P}(a_{jk}|c_i)=\frac{n_c+mp}{n+m}$$

(m-estimate)

n_c : number of training examples for which $x_j=a_{jk}$ and $c=c_i$

n : number of training examples for which $c=c_i$

p : prior estimate (usually, $p=1/t$ for t possible values of x_j)

m : weight to prior (number of "virtual" examples, $m\geq 1$)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Zero conditional probability

- Example: P(outlook=overcast|no)=0 in the play-tennis dataset
  - Adding ***m*** "virtual" examples (*m*: up to 1% of #training example)
    - In this dataset, # of training examples for the "no" class is 5.
    - We can only add ***m*=1** "virtual" example in our m-esitmate remedy.
  - The "outlook" feature can takes only 3 values. So ***p=1/3***.
  - Re-estimate P(outlook|no) with the m-estimate

$$\mathrm{P(overcast|no)}=\frac{0+1*\left(\frac{1}{3}\right)}{5+1}=\frac{1}{18}$$

$$\mathrm{P(sunny|no)}=\frac{3+1*\left(\frac{1}{3}\right)}{5+1}=\frac{5}{9} \qquad \mathrm{P(rain|no)}=\frac{2+1*\left(\frac{1}{3}\right)}{5+1}=\frac{7}{18}$$

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Mathematical formula of NBC

>How to calculate NBC for discrete features

>How to calculate NBC for continuous features