Machine Learning II

Prof. Dr. Tim Downie

Lecture 5 – 3rd November 2023

Conditional independence and the naive Bayes classifier

BHT Berliner Hochschule
für Technik

Contents

- Conditional independence
- Naive Bayes classifier
- Fitting a naive Bayes classifier
- Comparison to LDA
- Overview of Bayesian machine learning, Bayesian networks, Bayesian neural networks

Classification as conditional probability

A supervised machine learning method is called *regression* if the outcome variable is numeric, usually continuous. If the outcome is a nominal variable then the method is called *classification*, which is what we are considering today.

The outcome Y variable belongs to one of K classes: $Y \in \{1, 2, \ldots, K\}$. We also have predictor variables $x_1, \ldots, x_p$ which are known.

Suppose we know the probability that $Y{=}k$ given the values $x_1, \ldots, x_p$. This is written:

$$P(Y{=}k | x_1, \ldots, x_p) \tag{1}$$

A sensible classifier is one that chooses the value k which maximises this conditional probability.

This is called a **Bayes classifier** (last week) and in the language of Bayesian statistics Equation (1) is called the posterior probability of Y given $x_1, \ldots, x_p$

The Bayes classifier is a theoretical best case classifier, not an algorithm.

In practice we can rarely obtain this probability exactly, we need a method to approximate these probabilities, and there are many different methods to do this.

Lets take a step back and simplify the expression in (1) ...

Assume that there is only one predictor X and this is a discrete random variable.

The probability in Equation (1) becomes $P(Y{=}k|X{=}x)$

Bayes' Theorem "reverses" the direction of the conditioning, so the expression is in terms of the probability X given Y...

$$P(Y{=}k|X{=}x) = \frac{P(X{=}x|Y{=}k)p(Y{=}k)}{P(X{=}x)} \tag{2}$$

The denominator does not involve Y at all, so the value of k which maximises the left hand side also maximises $P(X{=}x|Y{=}k)P(Y{=}k)$

In other words if we compute

$$f(k) = P(X{=}x|Y{=}k)P(Y{=}k)$$

for every class k, then we can choose $\arg\max f(k)$ as the best class and

$$P(Y{=}k|X{=}x) = \frac{f(k)}{\sum_j f(j)}$$

In Bayesian statistics Equation (2) is often express as

$$P(Y{=}k|X{=}x) \propto P(X{=}x|Y{=}k)p(Y{=}k)$$

$P(X{=}x|Y{=}k)$ is called the likelihood function and $P(Y{=}k)$ the prior probability.

In many cases both the likelihood and prior are feasible to compute.

Now we consider when there are two discrete predictor variables X_1 and X_2.

Starting with (1) and applying Bayes' Theorem twice, we get

$$P(Y{=}k|X_1{=}x_1, X_2{=}x_2) \propto P(X_1{=}x_1|Y{=}k, X_2{=}x_2)P(X_2{=}x_2|Y{=}k)P(Y{=}k). \tag{3}$$

Proof on the blackboard

Independence and conditional independence

Revision:

X_1 and X_2 are independent when

$$P(X_1|X_2) = P(X_1)$$

"The probability of X_1 doesn't change if we know X_2"

Equivalently, X_1 and X_2 are independent when

$$P(X_1 \cap X_2) = P(X_1)P(X_2)$$

and

$$P(X_2|X_1) = P(X_2)$$

Example:

Rolling a six and you passing the course Machine Learning II are independent events.

Conditional independence

X_1 and X_2 are **conditionally independent given Y** when

$$P(X_1|Y, X_2) = P(X_1|Y)$$

"If we know Y then it's irrelevant what X_2 is!"

Equivalently, X_1 and X_2 are conditionally independent given Y when

$$P(X_1 \cap X_2|Y) = P(X_1|Y)P(X_2|Y)$$

and

$$P(X_2|Y, X_1) = P(X_2|Y)$$

Examples of conditional independence in practice

The event that a BHT student chosen at random this semester ...

studies *Statistical Computing*	X_1
studies *Computer Science for Big Data*	X_2
is a 1st semester Data Science Masters Student	Y

X_1 and X_2 are not independent. $P(X_1|X_2) > P(X_1)$.

They (probably) are independent once we know whether the student is a 1st Semester Data Science student:

$$P(X_1|Y, X_2) = P(X_1|Y)$$

(Y “causes” X_1 and Y “causes” X_2)

The event that a person ...

is a regular smoker	X_1
has high blood pressure	Y
suffers a heart attack	X_2

X_1 and X_2 are not independent but they are independent once we know that the person has high blood pressure or not,
(X_1 “causes” Y “causes” X_2)

In the *naive Bayes* (NB) model we assume that all of the predictor variables are conditionally independent given the class of Y.

Returning to Equation 3: if X_1 and X_2 are conditionally independent given Y

$$\begin{aligned} P(Y{=}k|X_1{=}x_1, X_2{=}x_2) &\propto P(X_1{=}x_1|Y{=}k, X_2{=}x_2)P(X_2{=}x_2|Y{=}k)P(Y{=}k) \\ &= \qquad P(X_1{=}x_1|Y{=}k)P(X_2{=}x_2|Y{=}k)P(Y{=}k) \end{aligned}$$

The classification for p variables, applying Bayes' Theorem p times, is :

$$\begin{aligned} P(Y{=}k|x_1, \dots, x_p) \propto P(x_1&|Y{=}k, x_2, \dots, x_p) \\ &\times P(x_2|Y{=}k, x_3, \dots, x_p) \\ &\cdots \times P(x_p|Y{=}k)P(Y{=}k) \end{aligned}$$

and **if** all of the predictor variables are conditionally independent,
once the class Y is known then each of the factors $P(x_j|Y{=}k, x_{j+1}, \dots, x_p)$ reduces to $P(x_j|Y{=}k)$; knowing the class tells us all we need to know about the distribution of x_j.

Under the Naive Bayes assumption

$$P(Y{=}k|x_1, \ldots, x_p) \propto P(Y{=}k) \prod_{j=1}^{p} P(x_j|Y{=}k)$$

Usually $P(x_j|Y{=}k)$ is much easier to estimate than $P(x_j|Y{=}k, x_{j+1}, \ldots, x_p)$.6

The conditional independence assumption of the Naive Bayes model is a strong assumption.

Is this a problem?

The conditional independence assumption of the Naive Bayes model is a strong assumption.

Is this a problem?

On a practical level we can argue “as long as the results are good, who cares if the predictor variables are conditionally independent on the class?”

The conditional independence assumption of the Naive Bayes model is a strong assumption.

Is this a problem?

On a practical level we can argue “as long as the results are good, who cares if the predictor variables are conditionally independent on the class?”

There is a more mathematical justification:
The conditional independence assumption may well be wrong, but the resulting difference in $P(Y{=}k|x_1, \ldots, x_p)$ using the assumption compared to not using the assumption, is not so large.

This difference is usually small enough, so the k which maximises the posterior probability is the same, whether or not the assumption is genuinely true. The resulting classifier will not be drastically different.

Fitting a naive Bayes classifier

$$P(Y{=}k|x_1,\ldots,x_p) \propto P(Y{=}k)\prod_{j=1}^{p} P(x_j|Y{=}k)$$

$P(Y{=}k)$ the prior probability is easy to estimate. Just obtain the relative frequencies for Y in the data, ignoring the values of the predictor variables.

$P(x_j|Y{=}k)$ is the *likelihood* of x_j given the class k. This factor will usually be fitted by assuming a distribution and estimating the parameters.

Nominal predictor variables

Suppose variable X_j takes values, which we can just relabel (for convenience) as “A”, “B”,“C” ,...

$P(X_j = A|Y{=}k)$ is estimated directly from the relative frequency of $Y{=}k$ from those observations where $X_j = A$, ignoring all the other $X_\bullet$ variables.

If the cross-tabulation of Y and X_j is:

X_j		A	B	C	D	Total
Y	Yes	26	27	7	4	64
	No	5	5	2	0	12
Total		31	32	9	4	76

The likelihood terms are found by taking the row relative frequencies:

X_j		A	B	C	D
Y	Yes	0.406	0.422	0.109	0.062
	No	0.417	0.417	0.167	0.000

So $P(X_j{=}B|Y{=}\text{Yes}) = 0.422$

Continuous predictor variables

A very common assumption is to use the normal distribution density for the likelihood

$$(x_j | Y{=}k) \sim N(\mu_{jk}, \sigma_{jk}^2)$$

This assumption makes the estimation of the parameters easy:

Estimate μ_{jk} by taking the mean of x_j for all training observations with label class k.

Estimate σ_{jk}^2 similarly using the observed variance of x_j in class k.

To avoid some combinations of j and k having very small joint frequencies, the assumption is often made that the variance $\sigma_{jk}^2 = \sigma_j^2$ is independent of the class.

Comments

- NB classifiers have been successful in text classification and spam filtering.
- NB classification can be directly used for adaptive learning, eg. as new emails come in.
- It is a fast algorithm.
- There is no means of directly testing whether a variable improves the model or not. Splitting the data into training and test data sets is recommended to assess the accuracy of the classifier. If a model with potentially many variables is to be fitted, it is best to test each variable using cross validation.

Simple Example: Titanic data

Classifying which of the crew and passengers on the Titanic survive based on: Class (or Crew), Sex and Age.

```
> NB_Titanic=naiveBayes(Survived ~., data=Titanic_df)
> NB_Titanic

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
      No      Yes
0.676965 0.323035

Conditional probabilities:
     Class
Y            1st        2nd        3rd       Crew
  No  0.08187919 0.11208054 0.35436242 0.45167785
  Yes 0.28551336 0.16596343 0.25035162 0.29817159
```

```
     Sex
Y           Male     Female
  No  0.91543624 0.08456376
  Yes 0.51617440 0.48382560

     Age
Y          Child      Adult
  No  0.03489933 0.96510067
  Yes 0.08016878 0.91983122
```

If we have no passenger information, we use the a-priori probabilities to predict the survival of a passenger:

$P(Y{=}\text{No}) = 0.677$ and $P(Y{=}\text{Yes}) = 0.323$

The predicted value is “No”.

If we know the passenger is a second class passenger

```
> TitanicPriors<-prop.table(NB_Titanic$apriori)
> TitanicPriors*NB_Titanic$tables$Class[,2]
Y
        No        Yes
0.07587460 0.05361199
```

$$P(Y = \text{No}|X_1 = 2) \propto P(Y = \text{No})P(X_1 = 2|Y = \text{No}) = 0.0759$$
$$P(Y = \text{Yes}|X_1 = 2) \propto P(Y = \text{Yes})P(X_1 = 2|Y = \text{Yes}) = 0.05369.$$

The predicted value is “No”

The actual posterior probabilities are

```
> prop.table(TitanicPriors*NB_Titanic$tables$Class[,2])
Y
       No       Yes
0.5859649 0.4140351
```

If we know the passenger is a second class passenger and a child, we can calculate the proportional posterior probabilities to choose the predicted value.

```
> TitanicPriors*NB_Titanic$tables$Class[,2]*NB_Titanic$tables$Age[,1]
Y
         No         Yes
0.002647973 0.004298008
```

$$
\begin{aligned}
P(Y = \text{No}|X_1 = 2) &\propto P(Y = \text{No})P(X_1 = 2|Y = \text{No})P(X_3 = 1|Y = \text{No}) &&= 0.00265 \\
P(Y = \text{Yes}|X_1 = 2) &\propto P(Y = \text{Yes})P(X_1 = 2|Y = \text{Yes})P(X_3 = 1|Y = \text{Yes}) &&= 0.00430
\end{aligned}
$$

The predicted value is “Yes”

The actual posterior probabilities are

```
> prop.table(TitanicPriors*NB_Titanic$tables$Class[,2]*
    NB_Titanic$tables$Age[,1])
Y
       No       Yes
0.3812237 0.6187763
```

Predictions for all passengers and the classification matrix.

```
> NB_Preds=predict(NB_Titanic,Titanic_df)
> table(Titanic_df$Survived,NB_Preds)
     NB_Preds
        No  Yes
  No  1364  126
  Yes  362  349
```

In the workshop you will repeat this analysis and obtain the sensitivity and specificity of this model.

Comparison between NB and linear discriminant analysis

The above naive Bayes model is very similar to the linear discriminant analysis (LDA) model from last week.

LDA always uses a multivariate normal distribution for the likelihood $P(x_1, \ldots, x_p | Y{=}k)$:

$$(x_1, \ldots, x_p | Y{=}k) \sim N(\boldsymbol{\mu}_k, \Sigma) \tag{LDA}$$

Σ is the Variance-Covariance, which allows the predictor variables to be correlated.

There is no assumption of conditional independence.

The predictor variables $x_1, \ldots, x_p$ have to be continuous because of the normally distributed likelihood.

In LDA the prior term $P(Y{=}k)$ is often implicitly chosen as $P(Y{=}k) = \frac{1}{K}$

Naive Bayes (NB) can use any distribution for the likelihood, but the correlation between *x*-variables must be zero, i.e. the variance-covariance matrix is diagonal.

LDA is can be thought of as a “non-naive” Bayes ie (i.e. without conditional independence) with a normal likelihood.

In both LDA and NB the common Variance-Covariance assumption between classes *k* can be relaxed, in discriminant analysis this results in the QDA model.

$$(x_1, \ldots, x_p | Y{=}k) \sim N(\boldsymbol{\mu}_k, \Sigma_k) \tag{QDA}$$