Naïve Bayes Classifier

Data Science & Analytics
University of Missouri

Classification Setting Revisited

- **Given a set of rows/instances/exampl es**
  - Each row is a set of cols/features
  - Learn a function
    - class = f(instance)
  - And classify
    - New instances (test dataset) into a class
- **Across domains there are different names for these components**

x_1	x_2	…	x_n	y
				y_1
				y2
				y2
				Y3
				y1

Alternative Terms for Rows

x_1	x_2	…	x_n

- **entities**
- **instances**
- **examples**
- **records**
- **transactions**
- **objects**
- **points**
- **feature-vectors**
- **tuples**

Alternative Terms for Columns

x_1	x_2	…	x_n

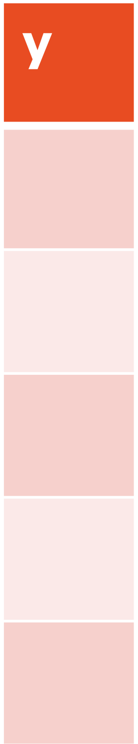
- **attributes**
- **properties**
- **features**
- **dimensions**
- **variables**
- **fields**
- **signal**

Alternative Terms for Input Variables

x_1	x_2	…	x_n

- independent variable
- predictor
- regressor (regression problem)
- covariate
- manipulated variable
- explanatory variable
- exposure variable (reliability theory)
- risk factor (see medical statistics),
- feature (in machine learning and pattern recognition)
- control variable (econometrics)
- exogenous (economics)

Alternative Terms for Output variable



- dependent variable
- response variable
- regressand (regression)
- criterion
- predicted variable
- measured variable
- explained variable
- experimental variable
- outcome variable
- target
- class
- label
- endogenous (economics)

Some Types of classifiers

- **Rule-based**
  - Decision tree (M1)
- **Probabilistic**
  - Naïve Bayes classifier (M2)
- **Max-margin classifier**
  - SVM (M5)
- **Neural network (M8)**

Applying NBC in practice

- **Classical example: text classification**
  - Instances are text samples
    - emails
    - paragraphs
    - sentences
    - documents
    - tweets
    - posts
    - comments
    - reviews
  - Classes are {spam, ~spam}

Applying NBC in practice

- **Real time prediction**
  - Fast to learn
- **Sentiment analysis**
- **Can be used in multi-class prediction**
  - Pr(Class|instance)
- **Need less training data**

Bayes Classifier

- **A probabilistic framework for solving classification problems**
- **Conditional Probability:** $$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

- **Bayes theorem:**

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Using Bayes Theorem for Classification

- **Consider each attribute and class label as random variables**
- **Given a record with attributes ($X_1, X_2, \ldots, X_d$), the goal is to predict class Y**
  - Specifically, we want to find the value of Y that maximizes $P(Y| X_1, X_2, \ldots, X_d)$
- **Can we estimate $P(Y| X_1, X_2, \ldots, X_d)$ directly from data?**

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

source: Tan et. Introduction to Data Mining

Bayesian Classifier: Using Bayes Theorem for Classification

- **Approach:**
  - compute posterior probability $P(Y \mid X_1, X_2, \ldots, X_d)$ using the Bayes theorem

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

  - *Maximum a-posteriori (MAP)*: Choose Y that maximizes

$$P(Y \mid X_1, X_2, \ldots, X_d)$$

    - MAP classification rule
  - Equivalent to choosing value of Y that maximizes

$$P(X_1, X_2, \ldots, X_d \mid Y)\, P(Y)$$

- **How to estimate $P(X_1, X_2, \ldots, X_d \mid Y)$?**

source: Tan et. Introduction to Data Mining

Two Types of Probabilistic Classification

- **Discriminative model**
  - Directly estimates the conditional class probability given the features

$$P(Y \mid X_1, X_2, \dots, X_d)$$

- **Generative model**
  - Estimates the conditional joint probability of features given the class variable

$$P(X_1, X_2, \dots, X_d \mid Y)$$

Generative vs Discriminative Classifiers

- **NBC is a generative classifier**
  - because it models $P(X_1, X_2, \ldots, X_n \mid c)$
  - Allows generate new instances by drawing samples the learned joint distribution
    - One can use the Naïve Bayes classifier to “generate” a document for a given class
- **A discriminative classifier, in contrast, will model P(class|instance), not P(instance| class)**
  - Example: logistic regression classifier, coming later

Why is Bayesian Classifier not Feasible in Practice?

- **Consider 3 classes, 10 features, each with 6 possible discrete instantiations**
  - We need $3*6^{10} + 3$ parameters to be estimated $\approx$181 million!

Enters Naïve Bayes

- **Conditional Independence Assumption**
  - **X** and **Y** are conditionally independent given **Z** if $P(\mathbf{X}|\mathbf{Y},\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})$
  - Example: Arm length and reading skills
    - Young child has shorter arm length and limited reading skills, compared to adults
    - If age is fixed, no apparent relationship between arm length and reading skills
    - Arm length and reading skills are conditionally independent given age

Disclaimer: By all accounts, Rev. Thomas Bayes was pretty smart. Naivete is on our part, not Bayes!

source: Tan et. Introduction to Data Mining

Naïve Bayes Classifier

- **Assume independence among attributes X_i when class is given:**
  - $P(X_1, X_2, \ldots, X_d | Y_j) = P(X_1 | Y_j)\, P(X_2 | Y_j) \ldots P(X_d | Y_j)$
  - Now we can estimate $P(X_i | Y_j)$ for all X_i and Y_j combinations from the training data
  - New point is classified to Y_j if $P(Y_j) \prod P(X_i | Y_j)$ is maximal.

source: Tan et. Introduction to Data Mining

Why is this a big deal?

- **Back to the Bayesian classifier: 3 classes, 10 features, each with 6 possible discrete instantiations**
  - Requires 181 million parameters:

- **Naïve Classifier:**
  - Requires 3*60 + 3 = 183 parameters!

source: Tan et. Introduction to Data Mining

Example Data

Given a Test Record: $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- We need to estimate

  P(Evade = Yes | X) and P(Evade = No | X)

  In the following we will replace

  Evade = Yes by **Yes**, and

  Evade = No by **No**

source: Tan et. Introduction to Data Mining

Example Data

Given a Test Record: $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Using Bayes Theorem:

- $P(\text{Yes} \mid X) = \dfrac{P(X \mid \text{Yes})P(\text{Yes})}{P(X)}$
- $P(\text{No} \mid X) = \dfrac{P(X \mid \text{No})P(\text{No})}{P(X)}$
- How to estimate $P(X \mid \text{Yes})$ and $P(X \mid \text{No})$?

source: Tan et. Introduction to Data Mining

Naïve Bayes on Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

P(X | Yes) =

P(Refund = No | Yes) x
P(Divorced | Yes) x
P(Income = 120K | Yes)

P(X | No) =

P(Refund = No | No) x
P(Divorced | No) x
P(Income = 120K | No)

source: Tan et. Introduction to Data Mining

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **P(y) = fraction of instances of class y**
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10
- **For categorical attributes:**

$$P(X_i = c|\, y) = n_c / \, n$$

  - where $|X_i = c|$ is number of instances having attribute value $X_i = c$ and belonging to class y
  - Examples:

    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

source: Tan et. Introduction to Data Mining

Estimate Probabilities from Data

- **For continuous attributes:**
  - Discretization: Partition the range into bins:
    - Replace continuous value with bin value
      - Attribute changed from continuous to ordinal
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

source: Tan et. Introduction to Data Mining

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **Normal distribution:**

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (X_i, Y_i) pair

- **For (Income, Class=No):**
  – If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

source: Tan et. Introduction to Data Mining

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(X | No) = P(Refund=No | No) $\times$ P(Divorced | No) $\times$ P(Income=120K | No) = 4/7 $\times$ 1/7 $\times$ 0.0072 = 0.0006
- P(X | Yes) = P(Refund=No | Yes) $\times$ P(Divorced | Yes) $\times$ P(Income=120K | Yes) = 1 $\times$ 1/3 $\times$ 1.2 $\times$ 10^{-9} = 4 $\times$ 10^{-10}

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Make Prediction with Partial Information in the Test Set

Even in absence of information about any attributes, we can use Apriori Probabilities of Class Variable:

P(Yes) = 3/10

P(No) = 7/10

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

If we only know that marital status is Divorced, then:

P(Yes | Divorced) = 1/3 x 3/10 / P(Divorced)

P(No | Divorced) = 1/7 x 7/10 / P(Divorced)

If we also know that Refund = No, then

P(Yes | Refund = No, Divorced) = 1 x 1/3 x 3/10 / P(Divorced, Refund = No)

P(No | Refund = No, Divorced) = 4/7 x 1/7 x 7/10 / P(Divorced, Refund = No)

If we also know that Taxable Income = 120, then

P(Yes | Refund = No, Divorced, Income = 120) = 1.2×10^{-9} x 1 x 1/3 x 3/10 / P(Divorced, Refund = No, Income = 120)

P(No | Refund = No, Divorced Income = 120) = 0.0072 x 4/7 x 1/7 x 7/10 / P(Divorced, Refund = No, Income = 120)

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

Naïve Bayes will not be able to classify X as Yes or No!

Issues with Naïve Bayes Classifier

- **If one of the conditional probabilities is zero, then the entire expression becomes zero**
- **Need to use other estimates of conditional probabilities than simple fractions**
- **Probability estimation:**

$$\text{original: } P(X_i = c|y) = \frac{n_c}{n}$$

$$\text{Laplace Estimate: } P(X_i = c|y) = \frac{n_c + 1}{n + v}$$

$$\text{m} - \text{estimate: } P(X_i = c|y) = \frac{n_c + mp}{n + m}$$

n: number of training instances belonging to class y

n_c: number of instances with $X_i = c$ and $Y = y$

v: total number of attribute values that X_i can take

p: initial estimate of $(P(X_i = c|y))$ known apriori

m: hyper-parameter for our confidence in p

Sklearn naïve Bayes

- https://scikit-learn.org/stable/modules/naive_bayes.html

naive_bayes.BernoulliNB(*[, alpha, ...])	Naive Bayes classifier for multivariate Bernoulli models.
naive_bayes.CategoricalNB(*[, alpha, ...])	Naive Bayes classifier for categorical features
naive_bayes.ComplementNB(*[, alpha, ...])	The Complement Naive Bayes classifier described in Rennie et al.
naive_bayes.GaussianNB(*[, priors, ...])	Gaussian Naive Bayes (GaussianNB)
naive_bayes.MultinomialNB(*[, alpha, ...])	Naive Bayes classifier for multinomial models

Conclusions

- **Naïve Bayes based on the independence assumption**
  - Training is very easy and fast; just requiring considering each attribute in each class separately
    - Works with less data
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- **A popular generative model**
  - Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  - Many successful applications, e.g., spam mail filtering
  - Apart from classification, naïve Bayes can do more…