Introduction to Machine Learning

Ranga Raju Vatsavai, Ph.D.
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)

Feb. 25-27, 2019

Probability - Basics

- Marginal Prob.
  - Prob. of an event occurring, p(A). It is not conditioned on another event.
    - Prob. that a card drawn from deck is black: 0.5

- Joint Prob.
  - Prob. of event A and B occurring, p(A and B). Its prob. of intersection of two or more events.
    - Prob. of card is black and 5 is: 1/26
- Conditional Probability
  - Prob. of event A occurring, given that event B occurs, p(A|B).
    - Say, you have drawn a black card, what's the prob. that's its 5. p(5|black) = 1/13

Probability - Basics

- Manipulation between conditional, joint, and marginal probabilities

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- Event A is independent of B, if P(A) = P(A|B)
  - Important consequence is multiplication rule

$$P(A \text{ and } B) = P(A \mid B)P(B) = P(A)P(B)$$

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

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Using Bayes Theorem for Classification

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

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

- Can we 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

Using Bayes Theorem for Classification

- Approach:
  - compute posterior probability P(Y | X_1, X_2, …, X_d) using the Bayes theorem

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

  - *Maximum a-posteriori*: Choose Y that maximizes P(Y | X_1, X_2, …, X_d)
  - Equivalent to choosing value of Y that maximizes P(X_1, X_2, …, X_d|Y) P(Y)
- How to estimate P(X_1, X_2, …, X_d | Y)?

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:

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

Naïve Bayes Classifier

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

Conditional Independence

- **X** and **Y** are conditionally independent given **Z** if **P(X|YZ) = P(X|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

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)

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

- Class: $P(Y) = N_c/N$
  - e.g., P(No) = 7/10, P(Yes) = 3/10
- For categorical attributes:

  $P(X_i \mid Y_k) = |X_{ik}|/N_{c_k}$
  - where $|X_{ik}|$ is number of instances having attribute value X_i and belonging to class Y_k
  - Examples:

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

Estimate Probabilities from Data

- For continuous attributes:
  - Discretization: Partition the range into K 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 $P(X_i|Y)$

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

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

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

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

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(Yes) = 3/10
  P(No) = 7/10
- P(Yes | Divorced) = 1/3 x 3/10 / P(Divorced)
  P(No | Divorced) = 1/7 x 7/10 / P(Divorced)
- 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)

Issues with Naïve Bayes Classifier

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(Yes) = 3/10
  P(No) = 7/10
- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

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
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

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:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability of the class

m: parameter

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

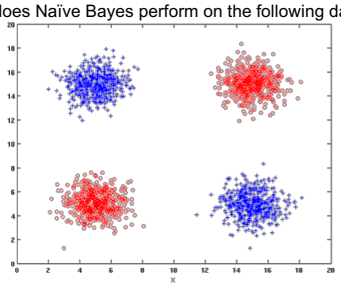
=> Mammals

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Naïve Bayes

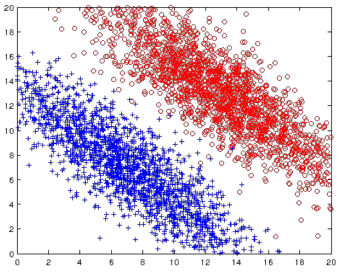
- How does Naïve Bayes perform on the following dataset?



Conditional independence of attributes is violated

Naïve Bayes

- How does Naïve Bayes perform on the following dataset?



Naïve Bayes can construct oblique decision boundaries

Maximum Likelihood Classification

- Uses Bayes decision rule:

$$x \in c_i \quad iff \quad p(c_i \mid x) > p(c_j \mid x) \forall j \neq i$$

How to compute $p(c_i \mid x)$?

Bayes' Theorem:

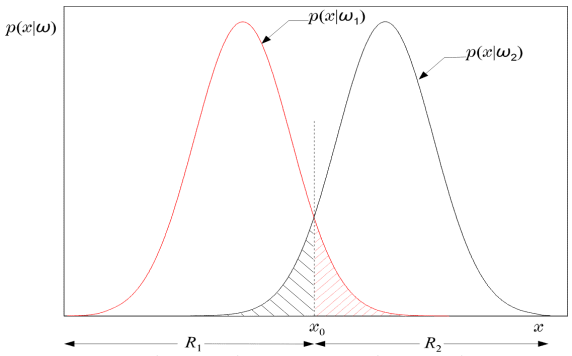
$$p(c_i \mid x) = \frac{p(x \mid c_i).p(c_i)}{p(x)}$$

Where $p(x)$ = prob. of finding a pixel from any class at location x

$$= \sum_{i=1}^{M} p(x \mid c_i).p(c_i).$$

Note: ω, y, c are used interchangeably for "class"

Decision Rule



$R_1(\rightarrow \omega_1)$ *and* $R_2(\rightarrow \omega_2)$

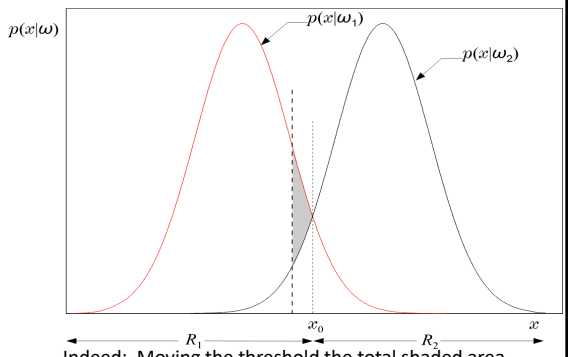
Error and Optimality

- Probability of error
  - Total shaded area

$$P_e = \int_{-\infty}^{x_0} p(x|\omega_2)dx + \int_{x_0}^{+\infty} p(x|\omega_1)dx$$

- Bayesian classifier is optimal with respect to minimizing the classification error probability

Optimality



Indeed: Moving the threshold the total shaded area INCREASES by the extra "grey" area.

MLC

$p(x \mid \omega_i)$: is the class conditional distribution and can be estimated from the training data (e.g., Gaussian)

$p(\omega_i)$: prior probability

Modeling both of these terms is very important for accurate classification

Decision Rule:

$$x \in \omega_i \text{ if } p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_j)p(\omega_j) \forall i \neq j$$

Simplified:

$$g_i(x) = \ln p(x \mid \omega_i) + \ln p(\omega_i)$$

$$x \in \omega_i \text{ if } g_i(x) > g_j(x) \forall i \neq j$$

MLC

- Model Assumption
  - Training samples were generated by multivariate Gaussian

$$p(x \mid \omega_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]$$

  - MLE $\hat{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} x_k; \hat{\Sigma}_i = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu}_1)(x_k - \hat{\mu}_1)'$
- For each feature vector x $p(\omega_i \mid x) = \frac{p(x \mid \omega_i).p(\omega_i)}{p(x)}$
- Apply, ML Decision

$$x \in \omega_i \text{ if } p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_j)p(\omega_j) \forall i \neq j$$

Discriminant function

- Ln(.) is monotonic. Define:

$$g_i(\underline{x}) = \ln(p(\underline{x}|\omega_i)P(\omega_i)) =$$

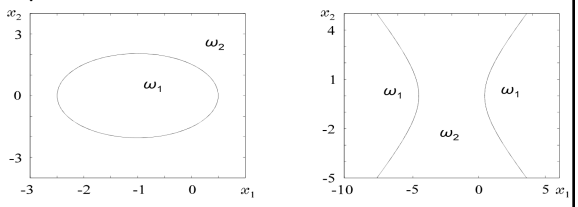
$$\ln p(\underline{x}|\omega_i) + \ln P(\omega_i)$$

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) + \ln P(\omega_i) + -(\frac{d}{2})\ln 2\pi - (\frac{1}{2})\ln|\Sigma_i|$$

Note: ω, y, c are used interchangeably for "class"

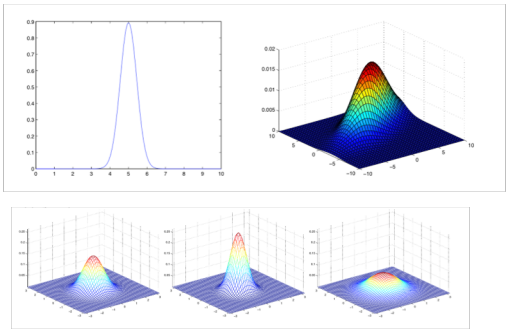
Discriminant function

- Note that $g_i(x)$ is quadratic, and produces quadrics, ellipsoids, parabolas, hyperbolas, and pairs of lines as decision boundaries

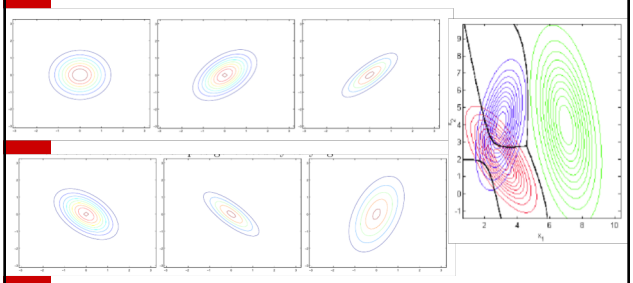


Accurate Assessment of Covariance is important, as it influences decision boundaries

Covariance Matrix



Covariance Matrix



Accurate Assessment of Covariance is important,
as it influences decision boundaries

Any Questions?