Lecture 9: Naive Bayes, SVM, Kernels

Instructor: Saravanan Thirumuruganathan

Outline

1. Probability basics
2. Probabilistic Interpretation of Classification
3. Bayesian Classifiers, Naive Bayes
4. SVM and Kernels

Probability Basics

Sample Space

- **Sample Space:** A space of events that we assign probabilities to
- Events can be binary, multi-valued, or continuous
- Events are mutually exclusive
- Examples:
  - Coin flip: $\{$head, tail$\}$
  - Die roll: $\{1,2,3,4,5,6\}$
  - English words: a dictionary
  - Temperature: $\mathbb{R}_+$ (Kelvin)

Random Variable

- A variable, X, whose domain is the sample space, and whose value is somewhat uncertain
- Examples:
  - $X =$ coin flip outcome
  - $X =$ first word in tomorrow's headline news
  - $X =$ tomorrow's temperature

Probability for Discrete Events

- Probability $P(X = a)$ is the fraction of times x takes value a
- Often we write it as P(a)
- Examples:
  - Fair Coin: P(head)=P(tail)=0.5
  - Slightly Biased Coin: P(head)=0.51, P(tail)=0.49
  - Two Face's Coin: P(head)=1, P(tail)=0
  - Fair Dice: P(getting 1 in a die roll) = 1/6

Probability for Discrete Events

- P(A="head or tail in a fair coin")

$$0.5 + 0.5 = 1$$

- P(A="even number in a fair dice roll")

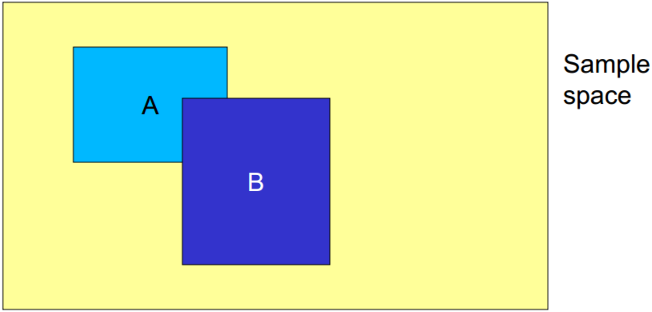
$$1/6 + 1/6 + 1/6 = 0.5$$

- P(A="two dice rolls sum to 2 in a fair dice")

$$1/6 * 1/6 = 1/36$$

Axioms of Probability

- $P(A) \in [0,1]$
- $P(true)=1$, $P(false)=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Simple Corollaries

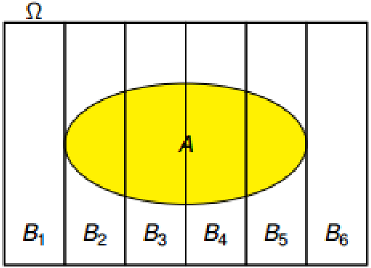
- $P(A') = 1 - P(A)$
- If A can take k different values $a_1, \ldots, a_k$,

$$P(A = a_1) + \ldots + P(A = a_k) = 1$$

- **Law of Total Probability:**
  - $P(A) = P(A \cap B) + P(A \cap B')$
  - $P(A) = \sum_{i=1}^{k} P(A \cap B = b_i)$ if B takes k values $b_1, \ldots, b_k$

Law of Total Probability[1]

Definition: A partition of the sample space Ω is a collection of disjoint events $B_1, B_2, \ldots B_k$ whose union is Ω. Such a partition divides any set A into disjoint pieces:



[1] http://people.reed.edu/~jones/Courses/P02.pdf

Probability Table

- Weather

Sunny	Cloudy	Rainy
200/365	100/365	65/365

- P(Weather = sunny) = P(sunny) = 200/365
- P(Weather) = {200/365, 100/365, 65/365}

Joint Probability Table

		weather	
temp	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

- P(temp=hot, weather=rainy) = P(hot, rainy) = 5/365
- The full joint probability table between N variables, each taking k values, has k^N entries (that's a lot!)

Marginal Probability Table

- Sum over other variables

weather

temp		Sunny	Cloudy	Rainy
	hot	150/365	40/365	5/365
	cold	50/365	60/365	60/365
	Σ	200/365	100/365	65/365

P(Weather)={200/365, 100/365, 65/365}

- The name comes from the old days when the sums are written on the margin of a page

Marginal Probability Table

- Sum over other variables

weather

temp		Sunny	Cloudy	Rainy	Σ
	hot	150/365	40/365	5/365	195/365
	cold	50/365	60/365	60/365	170/365

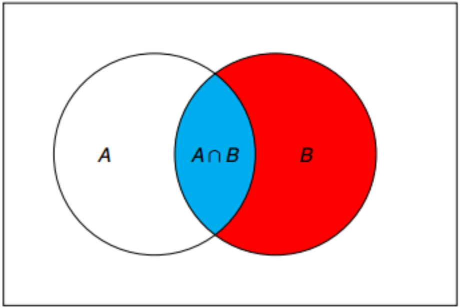
P(temp)={195/365, 170/365}

- This is nothing but $P(B) = \Sigma_{i=1...k} P(B \wedge A=a_i)$, if A can take *k* values

Conditional Probability

- $P(A = a|B = b)$ = fraction of times when random variable A took a value of a, within the region where random variable $B = b$

The definition $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ restricts the sample space to B, and rescales to give $\mathbb{P}(B|B) = 1$:



Conditional Probability

- Consider a roll of a fair dice
- A: it rolled 1. $P(A) = 1/6$
- B: it rolled an odd number. $P(B) = 3/6 = 0.5$
- Suppose, I knew that B happened. What is the probability that A happened?

Conditional Probability

- Consider a roll of a fair dice
- A: it rolled 1. $P(A) = 1/6$
- B: it rolled an odd number. $P(B) = 3/6 = 0.5$
- Suppose, I knew that B happened. What is the probability that A happened?
- $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

Conditional Probability

- **Conditional Probability:** $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- **Multiplication Rule:** $P(A \cap B) = P(A|B)P(B)$
- **Chain Rule:**
  - $P(A_1, A_2) = P(A_1)P(A_2|A_1)$
  - $$\begin{aligned} P(A_1, A_2, A_3) &= P(A_3|A_1, A_2)P(A_1, A_2) \\ &= P(A_3|A_1, A_2)P(A_2|A_1)P(A_1) \end{aligned}$$
  - $$\begin{aligned} P(A_1, A_2, A_3) &= P(A_1|A_2, A_3)P(A_2, A_3) \\ &= P(A_1|A_2, A_3)P(A_2|A_3)P(A_3) \end{aligned}$$
  - $P\left(\cap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P\left(A_i | \cap_{j=1}^{i-1} A_j\right)$

Bayes Theorem

- $P(A|B) = \frac{P(A \cap B)}{P(B})$
- $P(B|A) = \frac{P(A \cap B)}{P(A})$
- Proof of Bayes Theorem:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \\ P(A|B)P(B) &= P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

Independence

- Two events A, B are independent, if (all 3 definitions are equivalent)
  - $P(A \cap B) = P(A)P(B)$
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$

Independence Misused

A famous statistician would never travel by airplane, because he had studied air travel and estimated that the probability of there being a bomb on any given flight was one in a million, and he was not prepared to accept these odds.

One day, a colleague met him at a conference far from home. "How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:million, then the odds of two bombs are (1/1,000,000) x (1/1,000,000). This is a very, very small probability, which I can accept. So now I bring my own bomb along!"

An innocent old math joke

Independence

- Independence between random variables is typically obtained via domain knowledge
- Suppose A and B be two independent random variables that can take k different values $a_1, ldots, a_k$ and $b_1, \ldots, b_k$
- The joint probability table typically has k^2 parameters
- If random variables are independent, then only $2k - 2$ parameters
    - $k = 2$, 4 vs 2
    - $k = 10$, 100 vs 18
    - $k = 100$, $10,000$ vs 198
- This is something great for data mining!

Conditional Independence

- Random variables can be dependent, but **conditionally independent**
- Your house has an alarm
  - Neighbor John will call when he hears the alarm
  - Neighbor Mary will call when she hears the alarm
  - Assume John and Mary dont talk to each other
- JohnCall independent of MaryCall?

Conditional Independence

- Random variables can be dependent, but **conditionally independent**
- Your house has an alarm
  - Neighbor John will call when he hears the alarm
  - Neighbor Mary will call when she hears the alarm
  - Assume John and Mary dont talk to each other
- JohnCall independent of MaryCall?
  - No - If John called, likely the alarm went off, which increases the probability of Mary calling
  - $P(MaryCall|JohnCall) \neq P(MaryCall)$

Conditional Independence

- If we know the status of the alarm, JohnCall won't affect Mary at all
- $P(MaryCall|Alarm, JohnCall) = P(MaryCall|Alarm)$
- We say JohnCall and MaryCall are **conditionally independent**, given Alarm
- In general A, B are conditionally independent given C
  - $P(A|B, C) = P(A|C)$ or
  - $P(B|A, C) = P(B|C)$ or
  - $P(A, B|C) = P(A|C) * P(B|C)$

Probabilistic Interpretation of Classification

Probabilistic Classifiers

- Type of classifiers that, given an input, produces a *probability distribution* over a set of classes
  - Probability that this email is spam is X and not spam is Y
  - Probability that this person has tumour is X and no tumour is Y
  - Probability that this digit is 0 is X, 1 is Y, ...
- Most state of the art classifiers are probabilistic
- Even k-NN and Decision trees have probabilistic interpretations

Prior Probability

- $P(A)$: Prior or unconditional probability of A
- Your belief in A in the absence of additional information
- Uninformative priors
  - Principle of indifference: Assign equal probabilities to all possibilities
  - Coin toss, P(head)=P(tail)=1/2
- Often you get from domain knowledge or from data
- P(email is spam) = 0.8

Conditional Probability as Belief Update

- $P(A)$: Prior belief in A
- $P(A|B)$: Belief after obtaining information B
- $P(A|B, C)$: Belief after obtaining information B and C

Conditional Probability as Belief Update[2]

- Suppose you work as security guard in Airport
- Your job: look at people in security line and choose some for additional screening
- You want to pick passengers with high "risk"
- A: Passenger is high risk
- By experience, you know only 0.1% of passengers are high risk (Prior probability)

[2] http://www.quora.com/In-laymans-terms-how-does-Naive-Bayes-work

Conditional Probability as Belief Update[3]

- Consider a random person:
  - The probability that this person is high risk is A is 0.1%
  - Suppose you notice that the person is male

[3] http://www.quora.com/In-laymans-terms-how-does-Naive-Bayes-work

Conditional Probability as Belief Update[3]

- Consider a random person:
  - The probability that this person is high risk is A is 0.1%
  - Suppose you notice that the person is male
    - There are more male criminals than female ones
  - The passenger is nervous

[3] http://www.quora.com/In-laymans-terms-how-does-Naive-Bayes-work

Conditional Probability as Belief Update[3]

- Consider a random person:
  - The probability that this person is high risk is A is 0.1%
  - Suppose you notice that the person is male
    - There are more male criminals than female ones
  - The passenger is nervous
    - Most criminals are nervous but most normal passengers are not
  - The passenger is a kid

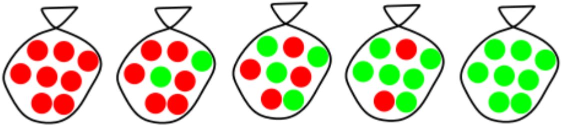
[3] http://www.quora.com/In-laymans-terms-how-does-Naive-Bayes-work

Conditional Probability as Belief Update[3]

- Consider a random person:
  - The probability that this person is high risk is A is 0.1%
  - Suppose you notice that the person is male
    - There are more male criminals than female ones
  - The passenger is nervous
    - Most criminals are nervous but most normal passengers are not
  - The passenger is a kid

[3] http://www.quora.com/In-laymans-terms-how-does-Naive-Bayes-work

Conditional Probability as Belief Update



Then we observe candies drawn from some bag: 

What kind of bag is it? What flavour will the next candy be?

Conditional Probability

- $X, A = \langle A_1, A_2, \ldots, A_d \rangle$: Input feature vector
- Y, C: Class value to predict
- $P(C|A)$ vs $P(A|C)$

Conditional Probability

- $X, A = \langle A_1, A_2, \dots, A_d \rangle$: Input feature vector
- Y, C: Class value to predict
- $P(C|A)$ vs $P(A|C)$
- Key terms
  - $P(A), P(C)$: Prior probability
  - $P(A|C)$: Class conditional probability or likelihood (from training data)
  - $P(C|A)$: Posterior probability

Likelihood vs Posterior

- $P(A|C)$: Likelihood, $P(C|A)$: Posterior
- Examples:
  - P(Viagra|Spam) and P(Spam|Viagra):

Likelihood vs Posterior

- $P(A|C)$: Likelihood, $P(C|A)$: Posterior
- Examples:
  - P(Viagra|Spam) and P(Spam|Viagra): Likelihood, Posterior
  - P(High temperature|Flu) and P(Flu|High Temperature):

Likelihood vs Posterior

- $P(A|C)$: Likelihood, $P(C|A)$: Posterior
- Examples:
  - P(Viagra|Spam) and P(Spam|Viagra): Likelihood, Posterior
  - P(High temperature|Flu) and P(Flu|High Temperature): Likelihood, Posterior
  - P(Fever, Headache, Cough|Flu) and P(Flu|Fever, Headache, Cough)
  - P(Viagra, Nigeria, Lottery|Spam) and P(Spam|Viagra, Nigeria, Lottery)

Bayes' Theorem

- Training data gives us likelihood and prior
- Prediction requires Posterior
- Bayes rule allows us to do statistical inference
- $P(C|A) = \frac{P(A|C)P(C)}{P(A)}$
- $P(C|A) \propto P(A|C)P(C)$
- Posterior $\propto$ Likelihood $\times$ Prior

Bayes Decision Rule

- "When you hear hoofbeats, think of horses not zebras"
- When predicting, assign the class with highest posterior probability
- To categorize email:
  - Compute P(spam|email) and P(not spam|email)
  - If P(spam|email) $>$ P(not spam|email), decide email as spam. Else as not-spam

Bayesian Classifiers

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

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?

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

Bayesian Classifiers

- Consider each attribute and class label as random variables
- Given a record with attributes $(A_1, A_2,\ldots,A_n)$
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes $P(C| A_1, A_2,\ldots,A_n)$
- Can we estimate $P(C| A_1, A_2,\ldots,A_n)$ directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n|C)\,P(C)$
- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naive Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j) P(A_2 | C_j) \ldots P(A_n | C_j)$
  - Can estimate $P(A_i | C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \Pi P(A_i | C_j)$ is maximal.

How to 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: $P(C) = N_c/N$
  - e.g., P(No) = 7/10, P(Yes) = 3/10
- For discrete attributes:

  $$P(A_i \mid C_k) = |A_{ik}| / N_{c_k}$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

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

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption k
  - Two-way split: $(A < v)$ or $(A > v)$
    - choose only one of the two splits as new attribute
  - 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, can use it to estimate the conditional probability $P(A_i|c)$

How to 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:

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

  - One for each (A_i, c_j) 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 Naive Bayes Classifier

Given a Test Record:

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

naive 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) = 2/7
P(Marital Status=Divorced|Yes)=1/7
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|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024
- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

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

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

Naive Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- 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

m: parameter

Example of Naive 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

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

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

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

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

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

Naive Bayes Classifier 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)

Support Vector Machines

Summary

Major Concepts:

- Probabilistic interpretation of Classification
- Bayesian Classifiers
- Naive Bayes Classifier
- Support Vector Machines (SVM)
- Kernels

Slide Material References

- Slides from ISLR book
- Slides from Jerry Zhu
- See also the footnotes