UVA CS 4501 - 001 / 6501 – 007 Introduction to Machine Learning and Data Mining

Lecture 13: Probability and Statistics Review (cont.) + Naïve Bayes Classifier

Yanjun Qi / Jane, PhD

University of Virginia
Department of
Computer Science

Announcements: Schedule Change

- Midterm – rescheduled
  - Oct. 30th / 3:35pm – 4:35pm
  - Homework 4 is totally for sample midterm questions
  - HW3 will be out next Monday, due on Oct 25th
  - HW4 will be out next Tuesday, due on Oct 29th (i.e. for a good preparation for midterm. Solution will be released before due time.)

  - Grading of HW1 will be available to you this evening
  - Solution of HW1 will be available this Wed
  - Grading of HW2 will be available to you next week
  - Solution of HW2 will be available next week

Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

Where are we ? ➔ Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

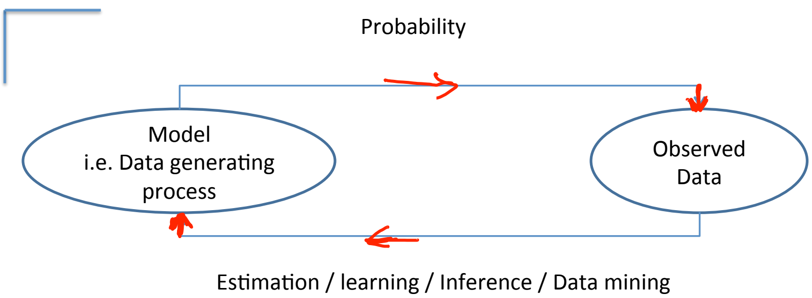
2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

Last Lecture Recap: Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability distributions, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Mean and Variance

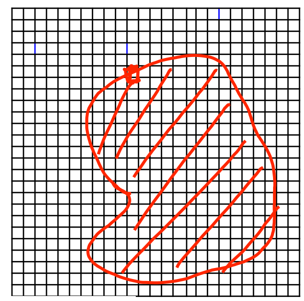
The Big Picture



But how to specify a model?

Probability as a measure of uncertainty

- *Probability is a measure of certainty of an event taking place.*
- *i.e. in the example, we were measuring the chances of hitting the shaded area.*



$$prob = \frac{\# \mathrm{Re}\,dBoxes}{\# Boxes}$$

Adapt from Prof. Nando de Freitas's review slides

e.g. Coin Flips

- You flip a coin
  - Head with probability 0.5
- You flip 100 coins
  - How many heads would you expect

e.g. Coin Flips cont.

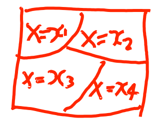
- You flip a coin
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*
- You flip *k* coins
  - How many heads would you expect
  - Number of heads X: discrete random variable
  - Binomial distribution with parameters *k* and *p*

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of heads X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2,…, 100

Probability of Discrete RV

- Probability mass function (pmf): P(X = x_i)
- Easy facts about pmf
  - Σ_i P(X = x_i) = 1
  - P(X = x_i ∩ X = x_j) = 0 if i ≠ j
  - P(X = x_i U X = x_j) = P(X = x_i) + P(X = x_j) if i ≠ j
  - P(X = x_1 U X = x_2 U … U X = x_k) = 1



Common Distributions

- Uniform $X \sim U[1,\ldots,N]$
  - X takes values 1, 2, …, N
  - $P(X = i) = 1/N$
  - E.g. picking balls of different colors from a box
- Binomial $X : Bin(n, p)$
  - X takes values 0, 1, …, n
  - $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$
  - E.g. coin flips

Coin Flips of Two Persons

- Your friend and you both flip coins
  - Head with probability 0.5
  - You flip 50 times; your friend flip 100 times
  - How many heads will both of you get

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together
  - E.g. P(You get 21 heads AND you friend get 70 heads)
-
  - E.g. $$\sum_x \sum_y P(X = x \cap Y = y) = 1$$

$$\sum_{i=0}^{50} \sum_{j=0}^{100} P(\text{You get } i \text{ heads AND your friend get } j \text{ heads}) = 1$$

Conditional Probability

- $P(X = x|Y = y)$ is the probability of $X = x$, given the occurrence of $Y = y$
  - E.g. you get 0 heads, given that your friend gets 61 heads
- $$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

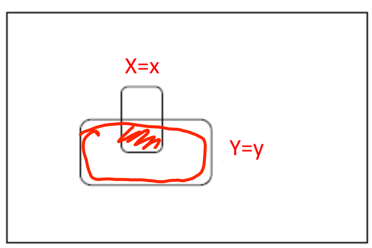
Conditional Probability

events

$$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

But we normally write it this way:

$$P(x|y) = \frac{p(x,y)}{p(y)}$$

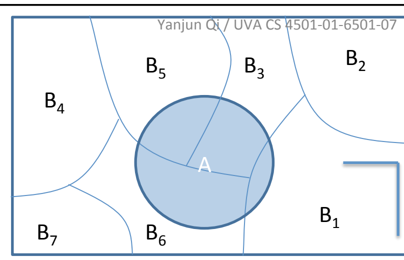


Law of Total Probability

- Given two discrete RVs X and Y, which take values in $\{x_1,\ldots,x_m\}$ and $\{y_1,\ldots,y_n\}$, We have

$$P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j)$$
$$= \sum_j P(X = x_i | Y = y_j) P(Y = y_j)$$

Marginalization



Marginal Probability

Joint Probability

$$P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j)$$

⇓ chain rule

$$= \sum_j P(X = x_i | Y = y_j) P(Y = y_j)$$

Conditional Probability

Marginal Probability

Bayes Rule

- X and Y are discrete RVs…

$$P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

$$P(X = x_i | Y = y_j) = \frac{P(Y = y_j | X = x_i) P(X = x_i)}{\sum_k P(Y = y_j | X = x_k) P(X = x_k)}$$

Today : Naïve Bayes Classifier

- ✓ Probability review
  - Structural properties, i.e., Independence, conditional independence
- ✓ Naïve Bayes Classifier
  - Spam email classification

Independent RVs

- Intuition: X and Y are independent means that $\mathrm{X} = x$ **neither** makes it **more or less** probable that $\mathrm{Y} = y$
- Definition: X and Y are independent ***iff***

$$\mathrm{P}(\mathrm{X} = x \cap \mathrm{Y} = y) = \mathrm{P}(\mathrm{X} = x)\mathrm{P}(\mathrm{Y} = y)$$

More on Independence

- $$\mathrm{P}(\mathrm{X} = x \cap \mathrm{Y} = y) = \mathrm{P}(\mathrm{X} = x)\mathrm{P}(\mathrm{Y} = y)$$

$$\mathrm{P}(\mathrm{X} = x | \mathrm{Y} = y) = \mathrm{P}(\mathrm{X} = x) \quad \mathrm{P}(\mathrm{Y} = y | \mathrm{X} = x) = \mathrm{P}(\mathrm{Y} = y)$$

- E.g. no matter how many heads you get, your friend will not be affected, and vice versa

More on Independence

- X is independent of Y means that knowing Y does not change our belief about X.
  - $P(X|Y=y) = P(X)$
  - $P(X=x, Y=y) = P(X=x) P(Y=y)$

  - The above should hold for all x_i, y_j
  - It is symmetric and written as $X \perp Y$

$X \perp Y$

Conditionally Independent RVs

- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z ***iff***

$$P(X = x \cap Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$$

$X \perp Y \mid Z$

More on Conditional Independence

$$P\left(X = x \cap Y = y \middle| Z = z\right) = P\left(X = x \middle| Z = z\right) P\left(Y = y \middle| Z = z\right)$$

$$P\left(X = x \middle| Y = y, Z = z\right) = P\left(X = x \middle| Z = z\right)$$

$$P\left(Y = y \middle| X = x, Z = z\right) = P\left(Y = y \middle| Z = z\right)$$

Today : Naïve Bayes Classifier

- ✓ Probability review
  - Structural properties, i.e., Independence, conditional independence
- ✓ Naïve Bayes Classifier
  - Spam email classification

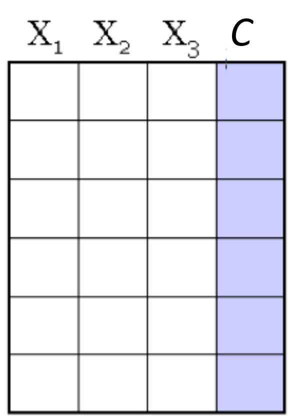
Where are we ? ➔ Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors



A Dataset for classification

$f : \boxed{X} \longrightarrow \boxed{C}$

Output as Discrete Class Label
$C_1, C_2, \ldots, C_L$

$\boxed{P(C \mid \mathbf{X})}$

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Bayes classifiers

- Treat each attribute and class label as random variables.
- Given a sample **x** with attributes (x_1, x_2, … , x_p):
  - Goal is to predict class *C.*
  - Specifically, we want to find the value of C_i that maximizes $p(C_i \mid x_1, x_2, \ldots , x_p)$.
- Can we estimate $p(C_i \mid \mathbf{x}) = p(C_i \mid x_1, x_2, \ldots , x_p)$ directly from data?

Bayes classifiers ➔ MAP classification rule

- Establishing a probabilistic model for classification

➔ **MAP** classification rule

- **MAP**: **M**aximum **A** **P**osterior
- Assign x to c^* if

$$\boxed{P(C=c^* \mid \mathbf{X}=\mathbf{x}) > P(C=c \mid \mathbf{X}=\mathbf{x})} \quad c \neq c^*, \; c = c_1, \cdots, c_L$$

Adapt from Prof. Ke Chen NB slides

Review: Bayesian Rule

- Prior, conditional and marginal probability
  - Prior probability: $P(C)$ $P(C_1), P(C_2), \ldots, P(C_L)$
  - Likelihood (through a generative model): $P(\mathbf{X} \mid C)$
  - Evidence (marginal prob. of sample): $P(\mathbf{X})$
  - Posterior probability: $P(C \mid \mathbf{X})$ $P(C_1|x), P(C_2|x), \ldots, P(C_L|x)$
- Bayesian Rule

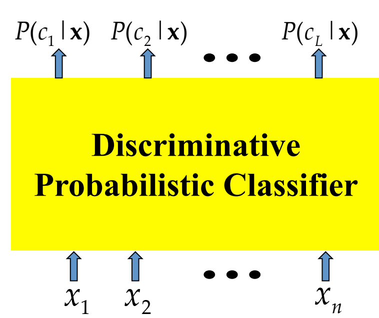
$$\boxed{P(C \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})}} \qquad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Adapt from Prof. Ke Chen NB slides

Bayes Classification Rule

- Establishing a probabilistic model for classification
  - **(1) Discriminative model**

$$P(C \mid \mathbf{X}) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$



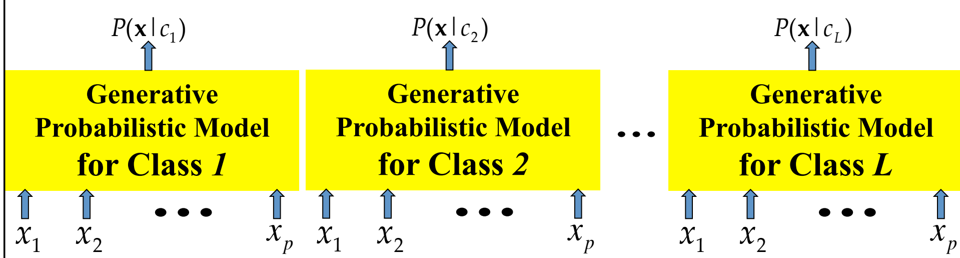
$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

Adapt from Prof. Ke Chen NB slides

Bayes Classification Rule

- Establishing a probabilistic model for classification (cont.)
  - **(2) Generative model**

$$P(\mathbf{X} / C) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_p)$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_p)$$

Adapt from Prof. Ke Chen NB slides

Yanjun Qi / UVA CS 4501-01-6501-07

Bayes Classification Rule

- MAP classification rule
  - **MAP**: **M**aximum **A** **P**osterior
  - Assign x to c^* if

$$P(C = c^* | \mathbf{X} = \mathbf{x}) > P(C = c | \mathbf{X} = \mathbf{x}) \quad c \neq c^*, \; c = c_1, \cdots, c_L$$

- Generative classification with the MAP rule
  - Apply Bayesian rule to convert them into posterior probabilities

$$P(C = c_i | \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} | C = c_i) P(C = c_i)}{P(\mathbf{X} = \mathbf{x})}$$
$$\propto P(\mathbf{X} = \mathbf{x} | C = c_i) P(C = c_i) \quad \text{for } i = 1, 2, \cdots, L$$

  - Then apply the MAP rule

Adapt from Prof. Ke Chen NB slides

Naïve Bayes

- Bayes classification

$$P(C \mid \mathbf{X}) \propto P(\mathbf{X} \mid C)P(C) = P(X_1, \cdots, X_p \mid C)P(C)$$

Difficulty: learning the joint probability $P(X_1, \cdots, X_p \mid C)$

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!

$$P(X_1, X_2, \cdots, X_p \mid C) = P(X_1 \mid X_2, \cdots, X_p, C)P(X_2, \cdots, X_p \mid C)$$
$$= P(X_1 \mid C)P(X_2, \cdots, X_p \mid C)$$
$$= P(X_1 \mid C)P(X_2 \mid C) \cdots P(X_p \mid C)$$

  - MAP classification rule: for $\mathbf{x} = (x_1, x_2, \cdots, x_n)$

$$[P(x_1 \mid c^*) \cdots P(x_p \mid c^*)]P(c^*) > [P(x_1 \mid c) \cdots P(x_p \mid c)]P(c),$$
$$c \neq c^*, \; c = c_1, \cdots, c_L$$

Adapt from Prof. Ke Chen NB slides

Yanjun Qi / UVA CS 4501-01-6501-07

Naïve Bayes

- Naïve Bayes Algorithm (for discrete input attributes)
  - Learning Phase: Given a training set S,

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

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

For every attribute value x_{jk} of each attribute X_j $(j = 1, \cdots, p;\; k = 1, \cdots, K_j)$

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

Output: conditional probability tables; for X_j, $K_j \times L$ elements

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

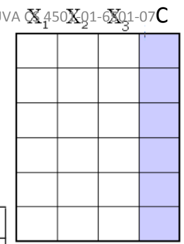
Look up tables to assign the label c^* to $\mathbf{X}'$ if

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

Adapt from Prof. Ke Chen NB slides

Example

- Example: Play Tennis



PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example

- Learning Phase

$P(X_2|C_1), P(X_2|C_2)$

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=*Ye* *s*	Play=*N* *o*
High	3/9	4/5
Normal	6/9	1/5

$P(X_4|C_1), P(X_4|C_2)$

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

$P(\text{Play}=Yes) = 9/14$ $\quad P(\text{Play}=No) = 5/14$ $\quad P(C_1), P(C_2), \ldots, P(C_L)$

Example

- Test Phase
  - Given a new instance,
    x' =(Outlook=*Sunny,* Temperature=*Cool,* Humidity=*High,* Wind=*Strong*)
  - Look up tables

P(Outlook=*Sunny*|Play=*Yes*) = 2/9	P(Outlook=*Sunny*|Play=*No*) = 3/5
P(Temperature=*Cool*|Play=*Yes*) = 3/9	P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*Yes*) = 3/9	P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*Yes*) = 3/9	P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*Yes*) = 9/14	P(Play=*No*) = 5/14

  - MAP rule

P(*Yes*|**x**'): [P(*Sunny*|*Yes*)P(*Cool*|*Yes*)P(*High*|*Yes*)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053

P(*No*|**x**'): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

Given the fact P(*Yes*|**x**') < P(*No*|**x**'), we label **x**' to be "*No*".

Adapt from Prof. Ke Chen NB slides

Next: Naïve Bayes Classifier

✓ Probability review
  - Structural properties, i.e., Independence, conditional independence

✓ Naïve Bayes Classifier
  - Text article classification

References

- Prof. Andrew Moore's review tutorial
- Prof. Ke Chen NB slides
- Prof. Carlos Guestrin recitation slides