UVA CS 6316 – Fall 2015 Graduate: Machine Learning

Lecture 12: Naïve Bayes Classifier

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

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

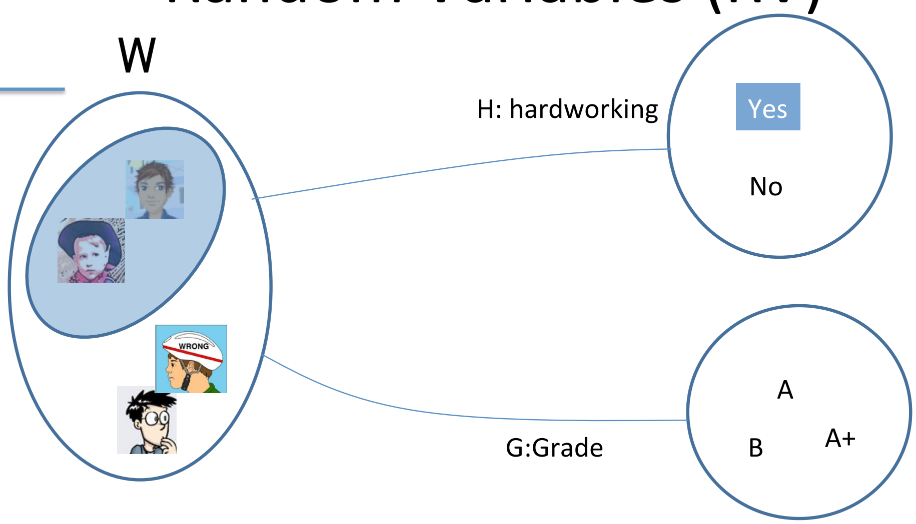
Last Lecture Recap: Probability Review

- The big picture : data <-> probabilistic model
- Sample space, Events and Event spaces
- Random variables
- Joint probability, Marginal probability, conditional probability,
- Chain rule, Bayes Rule, Law of total probability, etc.
- Structural properties
  - Independence, conditional independence

Sample space and Events

- W : Sample Space,
  - result of an experiment / set of all outcomes
  - If you toss a coin twice W= {HH,HT,TH,TT}
- Event: a subset of W
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Contains the empty event and W

Random Variables (RV)

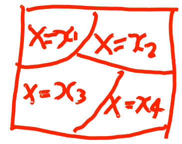


P(H = Yes) = P({all students who is working hard on the course})

- "functions" that maps from W to an attribute space T.

Probability of Discrete RV

- Probability mass function (pmf): $\boxed{P(X = x_i)}$
- Easy facts about pmf
  - $\Sigma_i\, P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$



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*

$$X \sim Bin(k, p) \qquad P(X = i) = \binom{k}{i} p^i (1-p)^{k-i}$$

Joint prob: e.g., Coin Flips by Two Persons

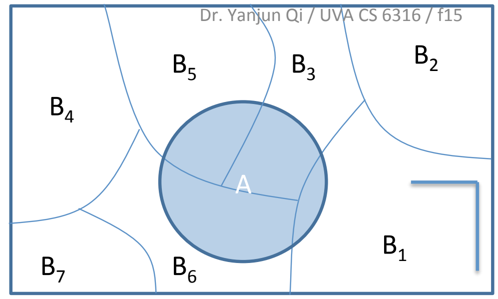
- Your friend and you both flip coins
  - Head with probability 0.5
  - You flip 50 times; your friend flip 100 times
  - How many heads will both of you get
- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together
  - E.g. P(You get 21 heads AND you friend get 70 heads)

$P((X=21) \wedge (y=70))$

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

Marginalization



Marginal Probability

Law of Total 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…

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

$$\Downarrow$$

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

Bayes Rule cont.

- You can condition on more variables

$$P(x \mid y, z) = \frac{P(x \mid z) P(y \mid x, z)}{P(y \mid z)}$$

$$p(x \mid y \mid z) = \frac{p(x, y \mid z)}{p(y \mid z)}$$

Conditional Probability Example

What is the probability that the 2nd ball drawn from the set {r,r,r,b} will be red?

Using marginalization, $P(B_2 = r)$ $= P(B_2 = r \wedge B_1 = r) + P(B_2 = r \wedge B_1 = b)$

$= P(B_1 = r) P(B_2 = r \mid B_1 = r) + P(B_1 = b) P(B_2 = r \mid B_1 = b)$

$$\begin{bmatrix} P(B_2=r) \\ P(B_2=b) \end{bmatrix} = \begin{bmatrix} P(B_2=r|B_1=r)P(B_1=r)+P(B_2=r|B_1=b)P(B_1=b) \\ P(B_2=b|B_1=r)P(B_1=r)+P(B_2=b|B_1=b)P(B_1=b) \end{bmatrix}$$

$$= \begin{bmatrix} P(B_2=r|B_1=r), & P(B_2=r|B_1=b) \\ P(B_2=b|B_1=r), & P(B_2=b|B_1=b) \end{bmatrix} \begin{bmatrix} P(B_1=r) \\ P(B_1=b) \end{bmatrix}$$

➔ Matrix Notation

$P(B_2)$ $\quad P(B_2|B_1)$ $\quad P(B_1)$

$\Rightarrow$ matrix notation form $\Pi_2 = G^T \Pi_1$

For short, we write this using vectors and a stochastic matrix:

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 $X = x$ **neither** makes it **more or less** probable that $Y = y$
- Definition: X and Y are independent ***iff***

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

$$p(X=x \mid Y=y)\,p(Y=y)$$

More on Independence

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

$$P(X = x \mid Y = y) = P(X = x) \qquad P(Y = y \mid X = x) = P(Y = y)$$

- 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***

X: sale ice-cream
Y: rate of Dro
Z: Weather

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

If holding for all x_i, y_j, z_k $\quad X \perp Y \mid Z$

More on Conditional Independence

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

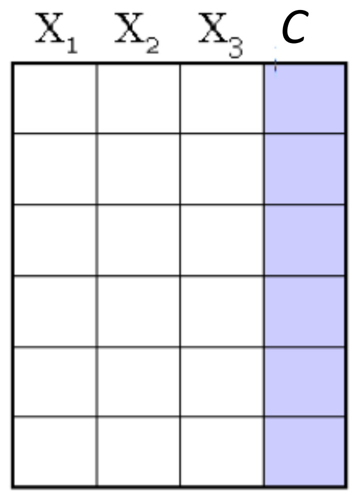
$$\frac{P(X=x, Y=y \mid Z=z)}{P(Y=y \mid Z=z)}$$

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z)$$

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

Today : Naïve Bayes Classifier

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



A Dataset for classification

$$f : X \longrightarrow C$$

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

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

$(x_1, x_2, \dots, x_p)$

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

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

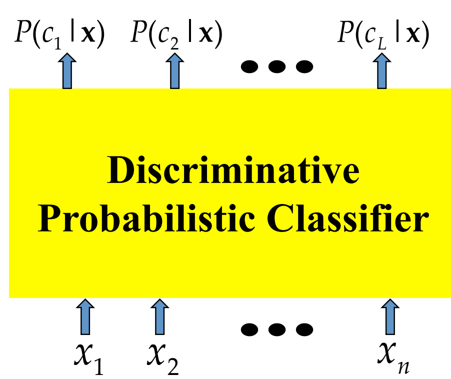
$$\left.\begin{cases} P(C=c_1 \mid x) \\ P(C=c_2 \mid x) \\ P(C=c_3 \mid x) \end{cases}\right\} \max \Rightarrow c_i$$

Adapt from Prof. Ke Chen NB slides

Bayes Classification Rule – (1)

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

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



logistic regression

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

Adapt from Prof. Ke Chen NB slides

Bayes Classification Rule – (2)

Bayes Rule

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

generative model

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

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

$P(\mathbf{x}\mid c_1)$	$P(\mathbf{x}\mid c_2)$		$P(\mathbf{x}\mid c_L)$
Generative Probabilistic Model for Class *1*	**Generative Probabilistic Model for Class *2***	•••	**Generative Probabilistic Model for Class *L***
x_1 x_2 ••• x_p	x_1 x_2 ••• x_p		x_1 x_2 ••• x_p

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

Adapt from Prof. Ke Chen NB slides

Review : Bayes' Rule – (2)

$$P(C,X) = P(C|X)P(X) = P(X|C)P(C)$$

$$\boxed{P(C|X) = \frac{P(X|C)P(C)}{P(X)}}$$

Prior

$P(C_1)$, $P(C_2)$, …, $P(C_L)$

Posterior

max ⇒ c^*

$\underbrace{P(C_1|x),\ P(C_2|x),\ \ldots,\ P(C_L|x)}$

$$P(C/\mathbf{X}) = \frac{P(\mathbf{X}/C)P(C)}{P(\mathbf{X})}$$

Review: Bayes Rule – (2)

- 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: $P(C \mid \mathbf{X})$ $\underbrace{P(C_1|x), P(C_2|x), \ldots, P(C_L|x)}$
- Bayes 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

(2) Generative classification with the MAP rule

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

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

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

$$P(C = c_i \mid \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_i)P(C = c_i)}{\cancel{P(\mathbf{X} = \mathbf{x})}}$$

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

$P(C=C_i)$

$P(X=x|C=C_i)$

  - Then apply the MAP rule

Adapt from Prof. Ke Chen NB slides

Naïve Bayes Classifier

$(X_1, X_2, \cdots, X_p)$

- Bayes classification

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

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

Difficulty: learning the joint probability

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

given class V

Naïve Bayes Classifier

$p(c)\,\underline{p(x|c)}$

- Naïve Bayes classification $\quad P(X_1, \cdots, X_p \mid C)$
  - Assumption that all input attributes are conditionally independent!

Chain Rule

$$\begin{aligned} P(X_1, X_2, \cdots, X_p \mid C) &= P(X_1 \mid X_2, \cdots, X_p, C)P(X_2, \cdots, X_p \mid C) \\ &= P(X_1 \mid C)P(X_2, \cdots, X_p \mid C) \\ &= P(X_1 \mid C)P(X_2 \mid C) \cdots P(X_p \mid C) \end{aligned}$$

CondInR

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

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

Naïve Bayes Classifier

- Naïve Bayes classification $P(X_1,\cdots,X_p \mid C)$
  - Assumption that all input attributes are conditionally independent!
  - MAP classification rule: for a testing sample

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

$P(X_j \mid c_i)$ $\{$ Bern, Bio, Multi, Gaussian

Naïve Bayes (for discrete input attributes) - training

- 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 **S**; → L

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 **S**;

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

$K_1, K_2, \ldots, K_p$

$\{X_1, X_2, \ldots, X_p\}$

$K_1 \times L +$
$K_2 \times L +$
$K_p \times L$

Naïve Bayes
(for discrete input attributes) - testing

- Naïve Bayes Algorithm (for discrete input attributes)
  - 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

$$[\hat{P}(a_1' \mid c^*) \cdots \hat{P}(a_p' \mid c^*)]\hat{P}(c^*) > [\underline{\hat{P}(a_1' \mid c) \cdots \hat{P}(a_p' \mid c)}]\hat{P}(c),$$

$$c \neq c^*, \ c = c_1, \cdots, c_L$$

$P(X' \mid C_i)\, P(C_i)$

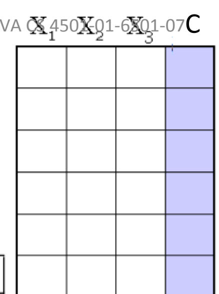
$= P(a_1' \mid C_i)\, P(a_2' \mid C_i) \ldots P(a_p' \mid C_i) P(C_i)$

$i = 1, 2, \cdots, L$

Example

$P(X_1, X_2, X_3, X_4 \mid Yes)$

$P(X_1, X_2, X_3, X_4 \mid No)$ $\rightarrow 36 \times 2$



- Example: Play Tennis

PlayTennis: training examples

3

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

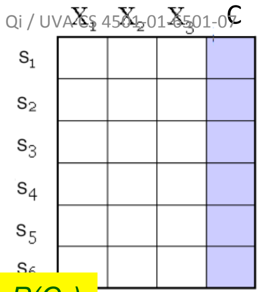
$3 \times 3 \times 2 \times 2 = 36$

$P(C = C_i)$

$\{Yes, No\}$

$P(C = Yes) = 9/14$

$P(C = No) = 5/14$



(a) Generative Bayes Classifier

$$P(C \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})}$$

- Learning Phase

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

$p=4$ $L; 2^p$

$P(X_1,X_2,\ldots,X_p|C_1), P(X_1,X_2,\ldots,X_p|C_2)$

Outlook (3 values)	Temperature (3 values)	Humidity (2 values)	X_4 Wind (2 values)	Play=*Yes*	Play=No
sunny	*hot*	*high*	*weak*	0/9	1/5
sunny	*hot*	*high*	*strong*	…/9	…/5
sunny	*hot*	*normal*	*weak*	…/9	…/5
sunny	*hot*	*normal*	*strong*	…/9	…/5
….	….	….	….	….	….
….	….	….	….	….	….
….	….	….	….	….	….
….	….	….	….	….	….

3*3*2*2 [conjunctions of attributes] * 2 [two classes]=72 parameters

(b) Naïve Bayes Classifier

$P(x_1, x_2, x_3, x_4 \mid C_i)$

Estimate $P(X_j = x_{jk} \mid C = c_i)$ with examples in **S**;

- Learning Phase

$P(X_2|C_1), P(X_2|C_2)$ 2

X_1 3

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

2

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

3

2

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

2

$P(X_4|C_1), P(X_4|C_2)$ 2

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

2

$\ll 72$

3+3+2+2 [naïve assumption] * 2 [two classes]= 20 parameters

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

$P(C_i)$ PL

(b) Naïve Bayes Classifier

$$[\hat{P}(a_1'|c^*)\cdots\hat{P}(a_p'|c^*)]\hat{P}(c^*) > [\hat{P}(a_1'|c)\cdots\hat{P}(a_p'|c)]\hat{P}(c)$$

- Test Phase
  - Given a new instance,

    x' =(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

$\rightarrow P(C_1)\,P(Sunny|C_1)\,P(Cool|C_1)\,P(High|C_1)\,P(Strong|C_1)$

$= \frac{9}{14} \times \frac{2}{9} \cdots\cdots =$

$\rightarrow P(C_2)\,P(Su|C_2)\,P(Co|C_2)\,P(hi|C_2)\,P(St|C_2)$

$= \frac{5}{14} \times \frac{3}{5} \times \cdots\cdots =$

(b) Naïve Bayes Classifier

$$[\hat{P}(a_1'|c^*)\cdots\hat{P}(a_p'|c^*)]\hat{P}(c^*) > [\hat{P}(a_1'|c)\cdots\hat{P}(a_p'|c)]\hat{P}(c)$$

- 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*".

Naïve Bayes Assumption

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_p/c_j)$
  - $O(|X|^p \cdot |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.

If no naïve assumption

$|C| \cdot p \cdot |X|$

Naïve Bayes Conditional Independence Assumption:

- Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i|c_j)$.

Adapt From Manning' textCat tutorial

References

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