Machine Learning

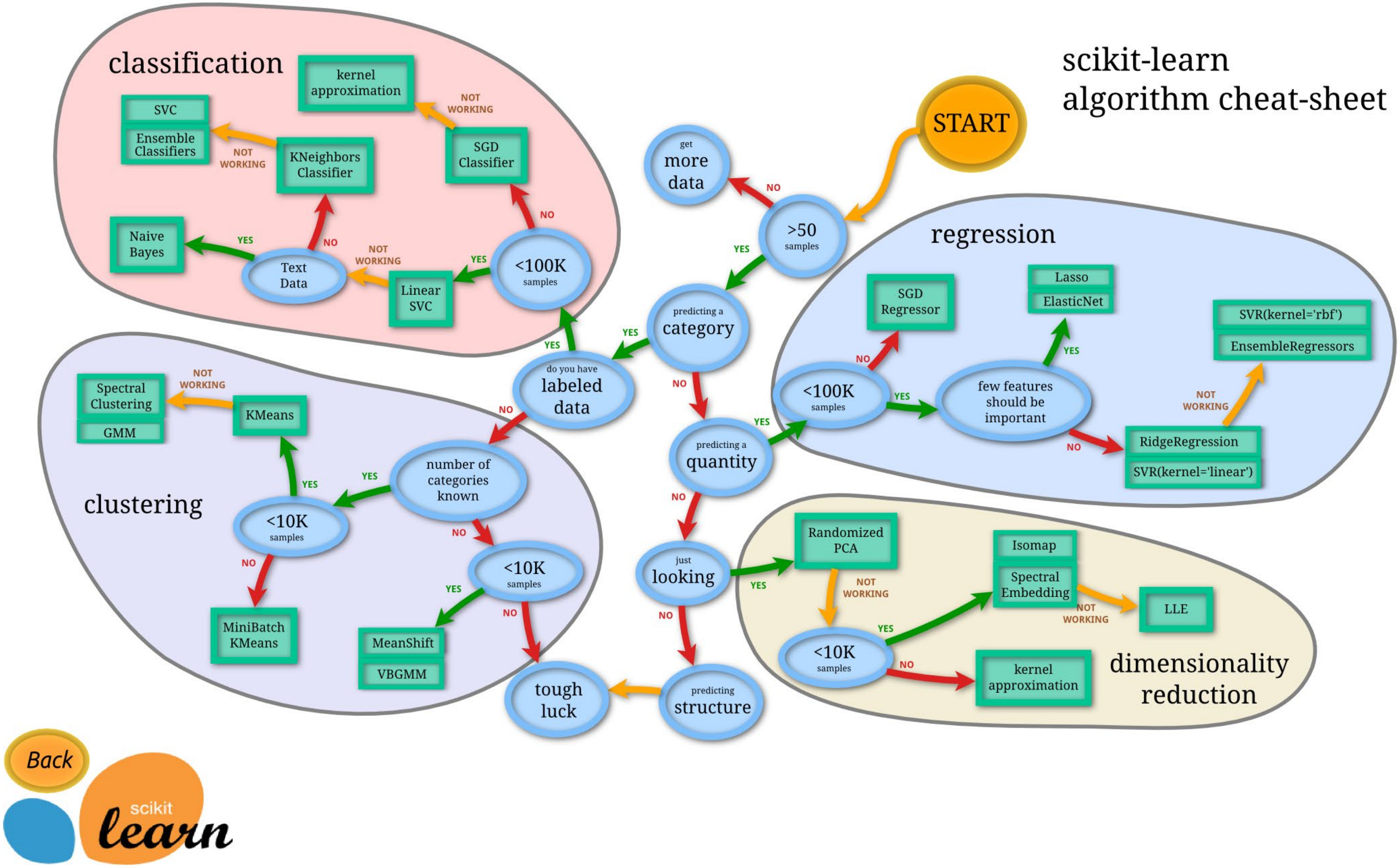
Lecture 17a: Generative Bayes Classifiers

Dr. Beilun Wang

Southeast University

School of Computer Science and Engineering

Roadmap



scikit-learn
algorithm cheat-sheet
START
classification
regression
clustering
dimensionality
reduction
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
<100K samples
Text Data
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Linear SVC
number of categories known
<10K samples
Spectral Clustering
GMM
KMeans
MiniBatch KMeans
MeanShift
VBGMM
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
few features should be important
RidgeRegression
SVR(kernel='linear')
Randomized PCA
Isomap
Spectral Embedding
LLE
YES
NO
NOT WORKING
Back
scikit learn

Course Content Plan

- ❑ Regression (supervised) — Y is a continuous
- ❑ Classification (supervised) — Y is a discrete
- ❑ Unsupervised models — NO Y
- ❑ Learning theory — About f()
- ❑ Graphical models — About interactions among X1,… Xp
- ❑ Reinforcement Learning — Learn program to Interact with its environment

Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types
  - Discriminative
    - directly estimate a decision rule/boundary
    - e.g., support vector machine, decision tree, logistic regression,
    - e.g. neural networks (NN), deep NN
  - Generative:
    - build a generative statistical model
    - e.g., Bayesian networks, Naïve Bayes classifier
  - Instance based classifiers
    - Use observation directly (no models)
    - e.g. K nearest neighbors

Today: Generative Bayes Classifiers

- Bayes Classifier (BC)
  - Generative Bayes Classifier
- Naïve Bayes Classifier
- Gaussian Bayes Classifiers
  - Gaussian distribution
  - Naïve Gaussian BC
  - Not-naïve Gaussian BC → LDA, QDA

Review: Bayes classifiers (BC)

- Treat each feature attribute and the class label as random variables.
- Testing: Given a sample $\boldsymbol{x}$ with attributes $(x_1, x_2, \dots, x_p)$:
  - Goal is to predict its class c.
  - Specifically, we want to find the class that maximizes $p(c|x_1, x_2, \dots, x_p)$
- Training: can we estimate $p(c_i|\boldsymbol{x}) = p(c_i|x_1, x_2, \dots, x_p)$ directly from data?

$$C_{MAP} = argmax_{c_i \in C} p(c_i|x_1, x_2, \dots, x_p)$$

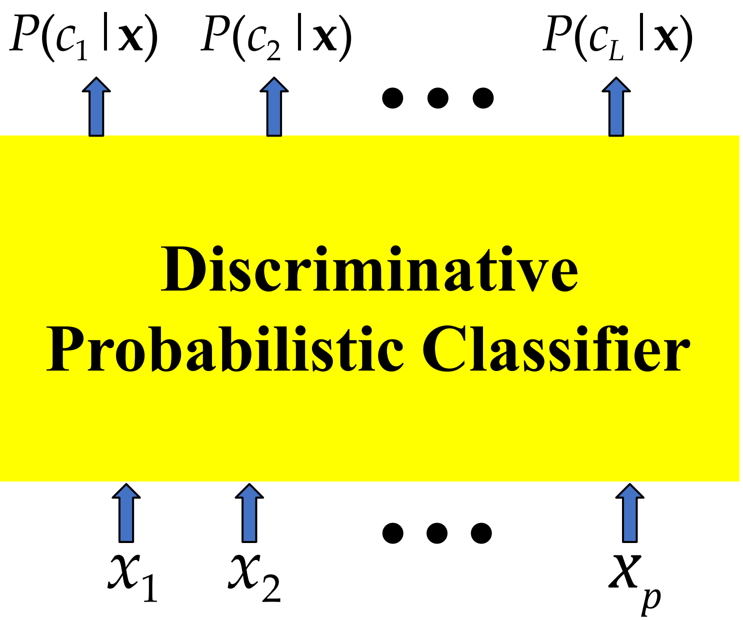
MAP Rule

Review: Two kinds of Bayes classifiers via MAP classification rule

- Establishing a probabilistic model for classification
  - Discriminative
  - Generative

Review: Discriminative BC

$$\underset{c \in C}{\operatorname{argmax}}\, P(c / \mathbf{X}), \quad C = \{c_1, \cdots, c_L\}$$



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

Review: Generative BC

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

$P(\mathbf{x}|c_1)$ $\Uparrow$ **Generative Probabilistic Model for Class *1*** $\Uparrow$ x_1 $\Uparrow$ x_2 $\bullet\bullet\bullet$ $\Uparrow$ x_p

$P(\mathbf{x}|c_2)$ $\Uparrow$ **Generative Probabilistic Model for Class *2*** $\Uparrow$ x_1 $\Uparrow$ x_2 $\bullet\bullet\bullet$ $\Uparrow$ x_p

$\cdots$

$P(\mathbf{x}|c_L)$ $\Uparrow$ **Generative Probabilistic Model for Class *L*** $\Uparrow$ x_1 $\Uparrow$ x_2 $\bullet\bullet\bullet$ $\Uparrow$ x_p

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

Review: Bayes Rule for Generative BC

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

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

$P(C_1), P(C_2), \ldots, P(C_L)$

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

Summary of Generative BC

- Apply Bayes rule to get posterior probabilities

$$P(C = c_i|\boldsymbol{X} = \boldsymbol{x}\,) = \frac{P(\boldsymbol{X} = \boldsymbol{x}\,|C = c_i)P(C = c_i)}{P(\boldsymbol{X} = \boldsymbol{x})}$$

$$\propto P(\boldsymbol{X} = \boldsymbol{x}\,|C = c_i)P(C = c_i)$$

$$for\ i = 1,2, \dots, L$$

- Then apply the MAP rule

Generative Bayes Classifier

Generative

Conditional Pr

Task	classification
↓	↓
Representation	Prob. models p(X|C)
↓	↓
Score Function	EPE with 0-1 loss → MAP Rule
↓	↓
Search/Optimization	Many options
↓	↓
Models, Parameters	Prob. Models' Parameter

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
	X_1	X_2	X_3	X_4	C

X_1	X_2	X_3	X_4	C
S (Sunny) O R	H (Hot) M C	H (High) N	W (Weak) S	$C_1 = Yes$ $C_2 = No$
3 ×	3 ×	2 ×	2 ×	2

→ 72 parameters

→ $P(X_1, X_2, X_3, X_4 | C), P(C)$

① $$\begin{cases} P(C = Yes) = \dfrac{N(Yes)}{N(train)} = \dfrac{9}{14} \\ P(C = No) = 1 - P(C = Yes) = \dfrac{5}{14} \end{cases}$$

② $$\begin{cases} P(X_1 = S, X_2 = H, X_3 = H, X_4 = W | C = No) = \dfrac{1}{5} \\ P(X_1 = S, X_2 = H, X_3 = H, X_4 = W | C = Yes) = 0 \end{cases}$$

Learning: maximum likelihood estimates

- simply use the frequencies in the data

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

e.g. $P(O, H, H, W|Yes) = \frac{1}{9}$

Directly estimate from data

e.g. $P(O, H, H, W|No) = \frac{0}{5} = 0$

Generative BC: Learning Phase

$P(C_1), P(C_2), \dots, P(C_L)$

$$P(Play = Yes) = 9/14 \qquad P(Play = No) = 5/14$$

$P(X_1, X_{2,\dots,} X_p | C_1), P(X_1, X_{2,\dots,} X_p | C_2)$

Outlook (3 values)	Temperature (3 values)	Humidity (2 values)	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

Generative BC: Testing Phase

- Given an unknown instance $\boldsymbol{x}'_{ts} = (a'_1, \dots, a'_p)$
  - Look up tables to assign the label c^* to $\boldsymbol{x}'_{ts}$ if

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

  - Given a new instance:

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

$$\begin{cases} P(x'|Yes)\, P(C = Yes) \\ P(x'|No)\, P(C = No) \end{cases} \Rightarrow \underset{C}{argmax} \Rightarrow predicted\ C^*$$

Today: Generative Bayes Classifiers

- Bayes Classifier (BC)
  - Generative Bayes Classifier
- Naïve Bayes Classifier
- Gaussian Bayes Classifiers
  - Gaussian distribution
  - Naïve Gaussian BC
  - Not-naïve Gaussian BC $\Rightarrow$ LDA, QDA

Naïve Bayes Classifier

Task
↓
Representation
↓
Score Function
↓
Search/Optimization
↓
Models, Parameters

classification
↓
$$P(X_1, X_2, \cdots, X_p | C) = P(X_1|C)P(X_2|C) \cdots P(X_p|C)$$

Prob. models p(X|C)
↓
EPE with 0-1 loss ➔ MAP Rule
↓
Many options
↓
Prob. Models' Parameter $P(X_j|C_i), P(C_i)$

Naïve Bayes Classifier

- Bayes classification

$$argmax_{c_j \in C} P(x_1, x_2, \ldots, x_p | c_j) P(c_j)$$

  - Difficulty: learning the joint probability
- Naïve Bayes classification
  - Assume that all input attributes are conditionally independent

Naïve Bayes Classifier

- Naïve Bayes classification
  - Assume that all input attributes are conditionally independent

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

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

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

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

Naïve Bayes Classifier (for discrete input attributes) – Training / Learning

– 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**;

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

Naïve Bayes Classifier (for discrete input attributes) – Testing

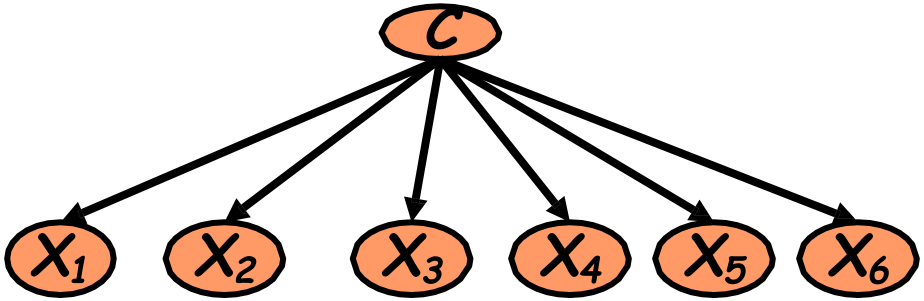
Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_p)$

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

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

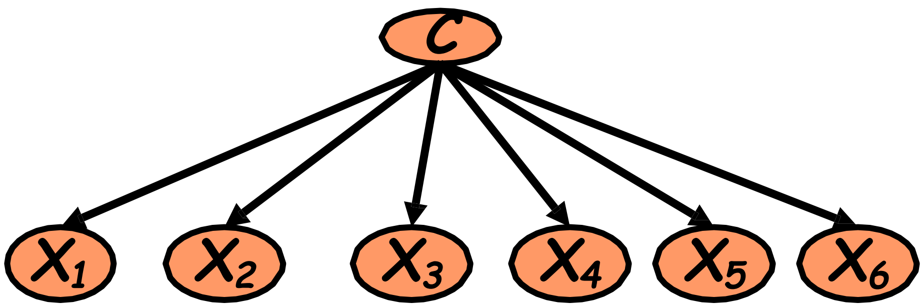
Learning (training) the NBC Model

$P(C_1) \cdots P(C_L)$



Learning (training) the NBC Model

$P(C_1) \cdots P(C_L)$



- maximum likelihood estimates:
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i|c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

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

$$P(X_1 = Rain|C = Yes) = \frac{3}{9}$$

$$P(X_1 = Rain|C = No) = \frac{2}{5}$$

$$\text{Estimate } P(X_j = x_{jk} | C = c_i) \text{ with examples in training;}$$

- Learning Phase

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

Outlook	Play=*Yes*	Play=*No*
Sunny		
Overcast		
Rain		

Temperature	Play=*Yes*	Play=*No*
Hot		
Mild		
Cool		

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

Humidity	Play=*Yes*	Play=No
High		
Normal		

Wind	Play=*Yes*	Play=*No*
Strong		
Weak		

P(Play=*Yes*) = ??

P(Play=*No*) = ??

$P(C_1), P(C_2), \ldots, P(C_L)$

Estimate $P(X_j = x_{jk} | C = c_i)$ with examples in training;

- 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

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

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

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

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

P(Play=*Yes*) = 9/14 *P*(Play=*No*) = 5/14

$P(C_1), P(C_2), \ldots, P(C_L)$

Example: Play tennis

- Test Phase

  Given a new instance,

  $\mathbf{x}'$=(Outlook=*Sunny,* Temperature=*Cool,* Humidity=*High,* Wind=*Strong*)

Example: Play tennis

- Test Phase

  Given a new instance,

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

  Look up in conditional-prob tables

P(Outlook=*Sunny*|Play=*Yes*) = 2/9

P(Temperature=*Cool*|Play=*Yes*) = 3/9

P(Huminity=*High*|Play=*Yes*) = 3/9

P(Wind=*Strong*|Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

P(Outlook=*Sunny*|Play=*No*) = 3/5

P(Temperature=*Cool*|Play==*No*) = 1/5

P(Huminity=*High*|Play=*No*) = 4/5

P(Wind=*Strong*|Play=*No*) = 3/5

P(Play=*No*) = 5/14

Testing the NBC Model

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*".

Why Naïve Bayes Assumption

Not Naïve

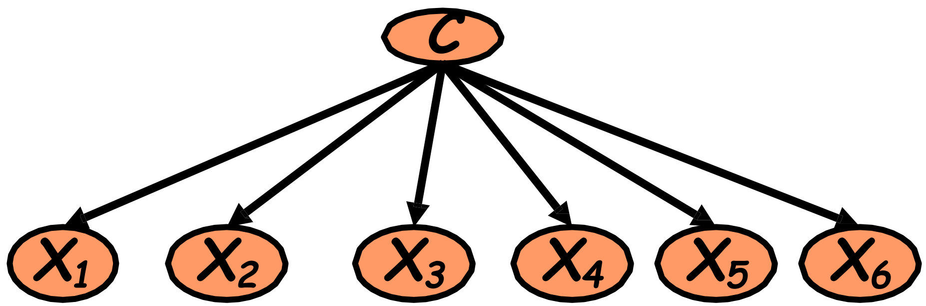
- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1, x_2, \ldots, x_p | c_j)$
  - $O(|X_1| \cdot |X_2| \cdot |X_3| \ldots |X_p| \cdot |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.

Naïve

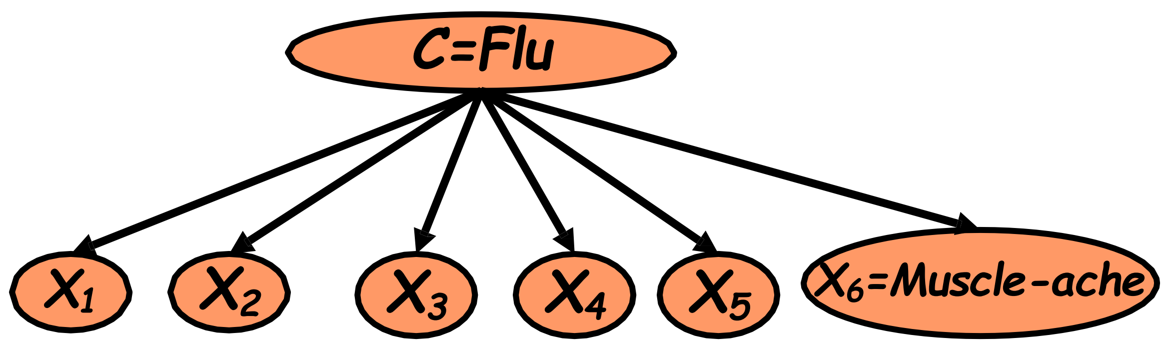
- $P(x_k | c_j)$
  - $O([|X_1| + |X_2| + |X_3| \ldots + |X_p|] \cdot |C|)$ parameters
  - Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i | c_j)$

Challenges during learning the NBC Model

$P(C_1) \cdots P(C_L)$



For instance:



Challenges during learning the NBC Model

- For instance:
  - What if we have seen no training cases where patient had no flu and muscle aches?

$$\hat{P}(X_6 = T|C = not\_flu) = \frac{N(X_6 = T, C = nf)}{N(C = nf)} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$?? = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i | c)$$

Smoothing to Avoid Overfitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k_i}$$

number of values of feature X_i

To make $sum_i(P(x_i|c_j)) = 1$

Smoothing to Avoid Overfitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k_i}$$

number of values of X_i

- Somewhat more subtle version

overall fraction in data where $X_i = x_{i,k}$

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + m p_{i,k}}{N(C = c_j) + m}$$

$k \in \{1, 2, \ldots, k_i\}$

extent of "smoothing"

Summary: Generative Bayes Classifier

- Task: Classify a new instance X based on a tuple of attribute values $X = \langle X_1, X_2, \ldots, X_p \rangle$ into one of the classes

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_p)$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_p \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_p)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_p \mid c_j) P(c_j)$$

MAP = Maximum A Posteriori

Next: Generative Bayes Classifiers

- Bayes Classifier (BC)
  - Generative Bayes Classifier
- Naïve Bayes Classifier
- Gaussian Bayes Classifiers
  - Gaussian distribution
  - Naïve Gaussian BC
  - Not-naïve Gaussian BC → LDA, QDA

References

- https://qiyanjun.github.io/2019f-UVA-CS6316-MachineLearning/
- Prof. Andrew Moore’s review tutorial
- Prof. Ke Chen NB slides
- Prof. Carlos Guestrin recitation slides

Thanks for listening