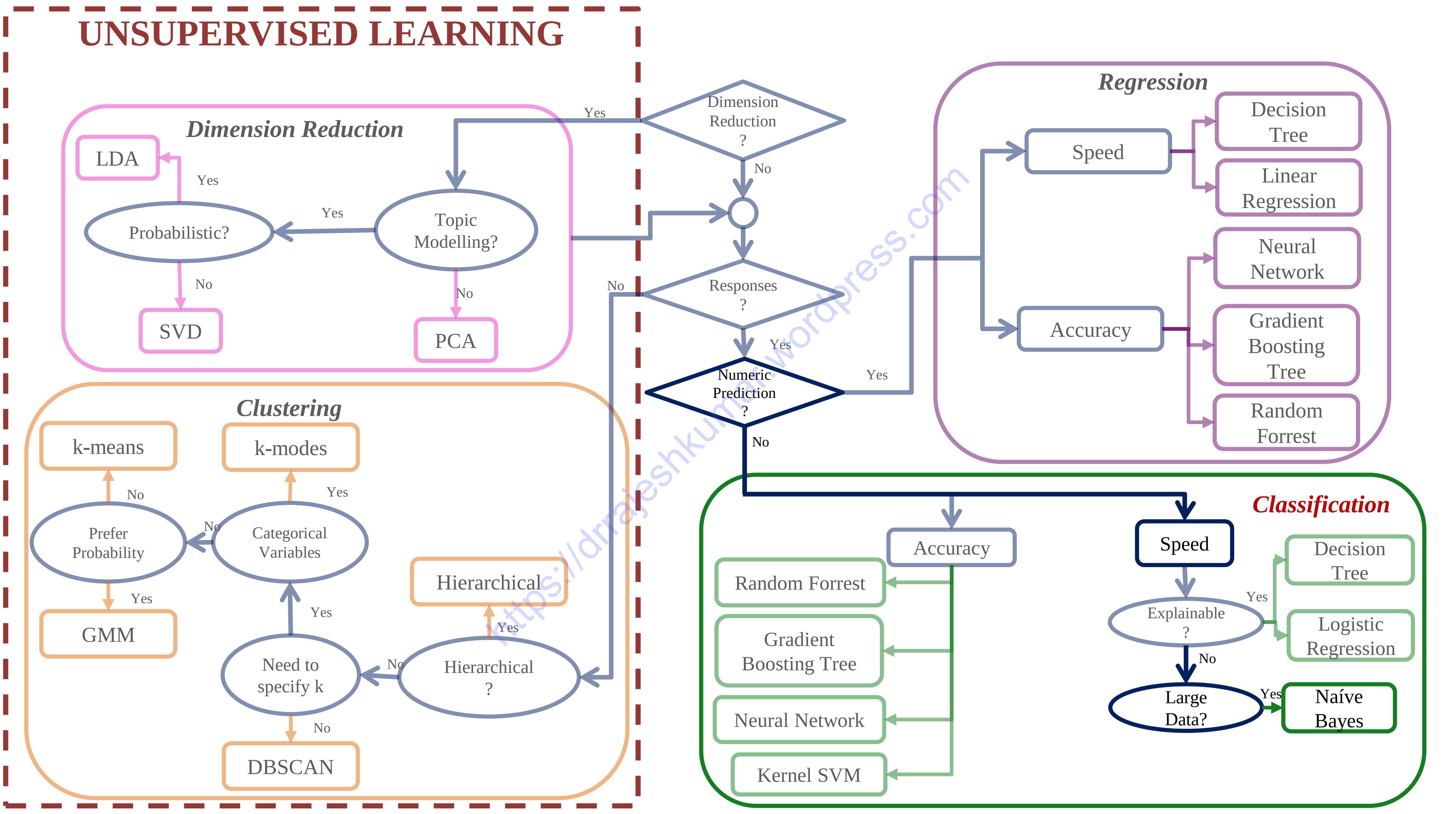
# Machine Learning

**Dr. Rajesh Kumar**

PhD, PDF (NUS, Singapore)
SMIEEE, FIET (UK), FIETE, FIE (I),SMIACSIT, LMISTE, MIAENG
Professor, Department of Electrical Engineering
Professor, Centre of Energy and Environment
Malaviya National Institute of Technology, Jaipur, India, 302017
Tel: (91) 9549654481
http://drrajeshkumar.wordpress.com
rkumar.ee@mnit.ac.in, rkumar.ee@gmail.com



UNSUPERVISED LEARNING
Dimension Reduction
LDA
Yes
Probabilistic?
Yes
Topic Modelling?
No
SVD
No
PCA
Clustering
k-means
k-modes
No
Yes
Prefer Probability
No
Categorical Variables
Hierarchical
Yes
GMM
Yes
Yes
Need to specify k
No
Hierarchical ?
No
DBSCAN
Dimension Reduction ?
Yes
No
Responses ?
No
Yes
Numeric Prediction ?
Yes
No
Regression
Speed
Decision Tree
Linear Regression
Accuracy
Neural Network
Gradient Boosting Tree
Random Forrest
Classification
Accuracy
Random Forrest
Gradient Boosting Tree
Neural Network
Kernel SVM
Speed
Explainable ?
Yes
Decision Tree
Logistic Regression
No
Large Data?
Yes
Naíve Bayes
https://drrajeshkumar.wordpress.com

# *Naive Bayes*

# Content

- Background
- Probability Basics
- Probabilistic Classification
- Naïve Bayes
  - Principle and Algorithms
  - Play Tennis
- Zero Conditional Probability
- Summary

# Probability Basics

- Prior, conditional and joint probability for random variables

  **Prior probability:** $P(x)$

  **Conditional probability:** $P(x_1|x_2), P(x_2|x_1)$

  **Joint probability:** $\mathbf{x} = (x_1, x_2),\ P(\mathbf{x}) = P(x_1, x_2)$

  **Relationship:** $P(x_1, x_2) = P(x_2|x_1)P(x_1) = P(x_1|x_2)P(x_2)$

  **Independence:** $P(x_2|x_1) = P(x_2),\ P(x_1|x_2) = P(x_1),\ P(x_1, x_2) =$

- Bayesian Rule

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

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Discriminative

Generative

# Probabilistic Classification

- **M**aximum **A** **P**osterior (**MAP**) classification rule
  - For an input $\boldsymbol{X}$, find the largest one from L probabilities output by a discriminative probabilistic classifier $P(c_1|\mathbf{x}), \ldots, P(c_L|\mathbf{x})$.
  - Assign $\boldsymbol{x}$ to label $c*$ if $P(c^*|\mathbf{x})$ is the largest.

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

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})} \propto P(\mathbf{x}|c_i)P(c_i)$$

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

Common factor for all $L$ probabilities

  - Then apply the MAP rule to assign a label

# Spam Detector

"Buy" and "Cheap"

Spam

No spam

# Spam Detector

"Buy" and "Cheap"

Spam

100%

No spam

0%

**Quiz:** If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

- ○ 40%
- ○ 60%
- ○ 80%
- ● 100%

# Solution: Collect more data?

# Spam Detector

"Buy" and "Cheap"

Spam

No spam


12 e-mails

0 e-mails?

# Spam Detector

Spam

25 e-mails
20 “Buy”
15 Cheap

4/5, 3/5 → 12/25 × 25 = 12 “Buy” and “Cheap”

# Spam Detector

No spam

75 e-mails

5 “Buy” 1/15

10 “Cheap” 2/15

2/225

# Spam Detector

No spam

75 e-mails

5 "Buy" 1/15

10 "Cheap" 2/15

$\rightarrow$ 2/225 × 75 = 2/3 "Buy" and "Cheap"

# Spam Detector

"Buy" and "Cheap"

**Quiz:** If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

Spam

$\underbrace{12}_{94.737\%}$

No spam

$\underbrace{2/3}_{5.263\%}$

$$\frac{12}{12 + 2/3} = \frac{36}{38}$$

$$= 94.737\%$$

# Naive Bayes

| | Spam | | No spam | |
|---|---|---|---|---|
| Total | 25 | | 75 | |
| Buy | 20 | 4/5 | 5 | 1/15 |
| Cheap | 15 | 3/5 | 10 | 2/15 |
| Buy & Cheap | 12 | 12/25 | 2/3 | 2/225 |

$$\frac{12}{12 + 2/3} = \frac{36}{38}$$

# Naive Bayes

| | Spam | | No Spam | |
|---|---|---|---|---|
| Total | 25 | | 75 | |
| Buy | 20 | 4/5 | 5 | 1/15 |
| Cheap | 15 | 3/5 | 10 | 2/15 |
| Work | 5 | 1/5 | 30 | 6/15 |
| Buy, Cheap, & Work | 12/5 | 12/125 | 4/15 | 12/3375 |

$$\frac{12/5}{12/5 + 4/15} = \frac{36}{40} = 90\%$$

# Bayes Theorem

S: Spam
H: Ham (not spam)
B: 'Buy'

$$P(S \mid B) = \frac{P(B \mid S)\,P(S)}{P(B \mid S)\,P(S) + P(B \mid H)\,P(H)}$$

$$P(\text{spam if "Buy"}) = \frac{\frac{20}{25}\;\frac{25}{100}}{\frac{20}{25}\;\frac{25}{100} + \frac{5}{75}\;\frac{75}{100}} = 80\%$$

# Naive Bayes

S: Spam
H: Ham (not spam)
B: 'Buy'
C: 'Cheap'

$$P(S \mid B \cap C) = \frac{P(B \mid S)\,P(C \mid S)\,P(S)}{P(B \mid S)\,P(C \mid S)\,P(S) + P(B \mid H)\,P(C \mid H)\,P(H)}$$

$$P(\text{spam if "Buy" \& "Cheap"}) = \frac{\frac{20}{25}\,\frac{15}{25}\,\frac{25}{100}}{\frac{20}{25}\,\frac{15}{25}\,\frac{25}{100} + \frac{5}{75}\,\frac{10}{75}\,\frac{75}{100}}$$

$$= 94.737\%$$

# 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: Play Tennis

- Learning Phase

| 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=*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 |

$P(\text{Play}=Yes) = 9/14$ $\quad$ $P(\text{Play}=No) = 5/14$

# Example: Play Tennis

- Test Phase
  - Given a new instance, predict its label

    **x**'=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables achieved in the learning phrase

| | |
|---|---|
| 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 |

  - Decision making with the 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*".

# Naive Bayes

- Example: Continuous-valued Features
  - Temperature is naturally of continuous value.

**Yes**: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

**No**: 27.3, 30.1, 17.4, 29.5, 15.1

  - Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

  - **Learning Phase**: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2\times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2\times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$