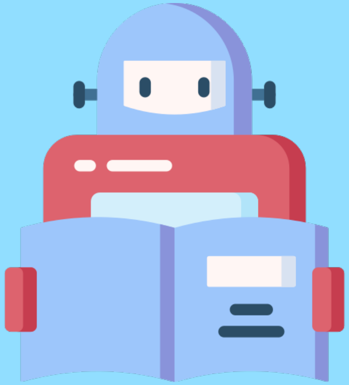
Mathematics for Machine Learning

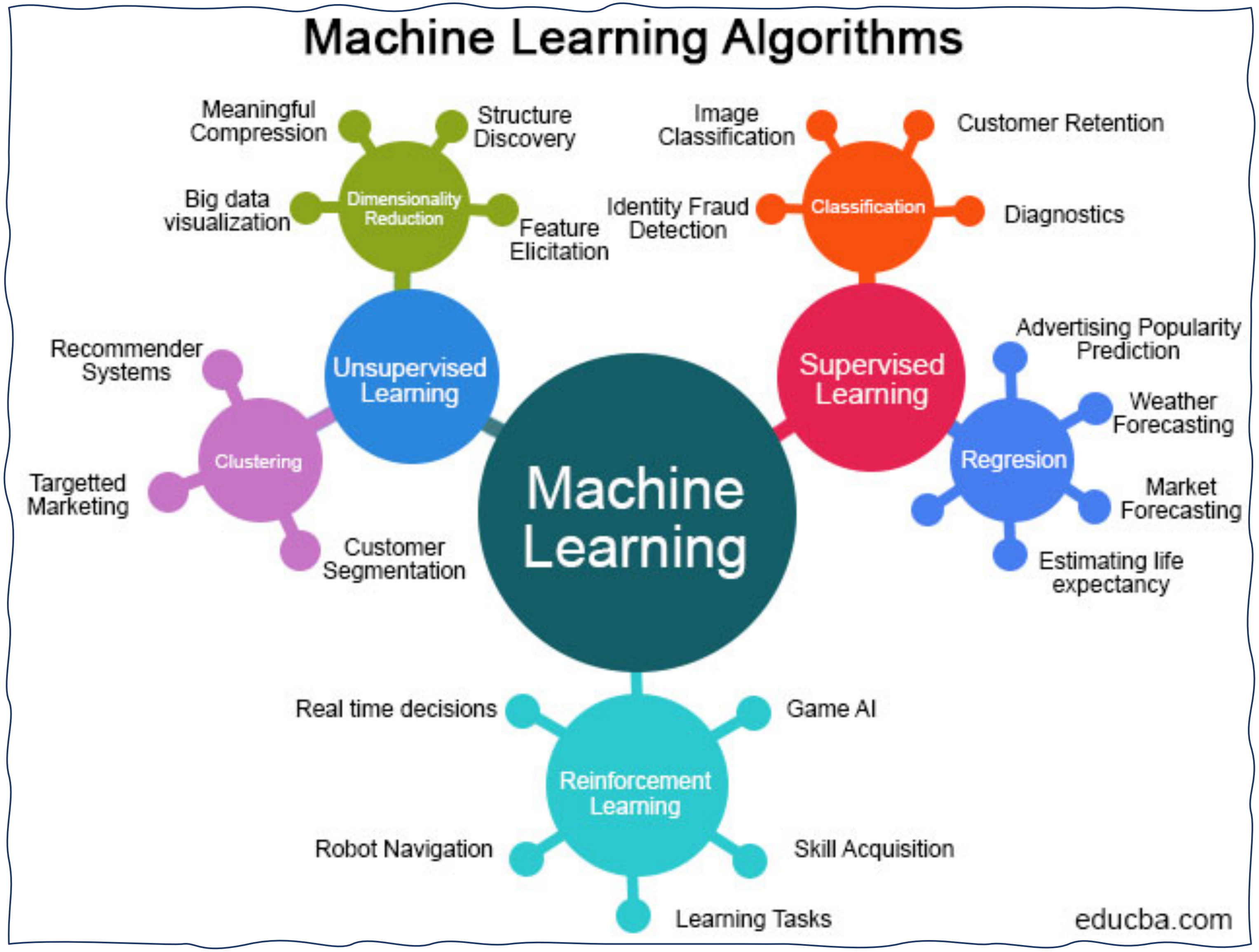


Special Mathematics Lecture
Nagoya University, Fall 2020

Lecture 5
Generative Learning algorithms & Naive Bayes

https://www.henrikbachmann.com/mml_2020.html

•



Machine Learning Algorithms
Meaningful Compression
Structure Discovery
Big data visualization
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Identity Fraud Detection
Classification
Diagnostics
Recommender Systems
Unsupervised Learning
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Clustering
Targetted Marketing
Regresion
Machine Learning
Market Forecasting
Customer Segmentation
Estimating life expectancy
Real time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks
educba.com

Generative vs Discriminative learning algorithm

Notation: P(A|B) refers to the conditional probability that event A occurs, given that event B has occurred.

x: feature (e.g. hours of studying)
y: label (e.g. passing or failing exam)

Logistic regression

Want to find a hypothesis which describes P(y|x)

$$P(y = 1 \mid x; \theta) = h_\theta(x) \, .$$

Learning P(y|x) is an example of a **discriminative learning algorithm.**

Generative learning algorithm: Learn P(x | y) and P(y).

Generative vs Discriminative learning algorithm

Notation: P(A|B) refers to the conditional probability that event A occurs, given that event B has occurred.

x: feature (e.g. hours of studying)
y: label (e.g. passing or failing exam)

Generative learning algorithm: Learn P(x | y) and P(y).

Question: But we want P(y | x)… don't we??

Generative vs Discriminative learning algorithm

Notation: P(A|B) refers to the conditional probability that event A occurs, given that event B has occurred.

x: feature (e.g. hours of studying)
y: label (e.g. passing or failing exam)

Generative learning algorithm: Learn P(x | y) and P(y).

Question: But we want P(y | x)… don't we??

Yes, but we can use:

Bayes rule:

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

Bayes rule

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

Generative learning algorithm

x: feature
y: label

Generative learning algorithm: Learn $P(x \mid y)$ and $P(y)$.

How does this help?

Naive Bayes

An example for a generative learning algorithm: Naive Bayes

Example: Spam filter

- Feature: Email
- Label: Spam & No Spam

$\mathcal{X} =$

$\mathcal{Y} = \{0, 1\}$

Naive Bayes assumption

Naive Bayes assumption:

The features are “conditionally independent” given the label.

Naive Bayes assumption

Naive Bayes assumption:

The features are “conditionally independent” given the label.

conditionally independent

A and B are conditionally independent given C if and only if, given knowledge that C occurs, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.

Naive Bayes assumption

Naive Bayes assumption:

The features are “conditionally independent” given the label.

If x_1, x_2 are conditionally independent given y, then we have

$$P(x_1 \mid y, x_2) = P(x_1|y)\,.$$

We want to calculate $P(x|y) = P(x_1, \ldots, x_d|y)$.

Chain rule of probabilities: $P(A, B) = P(A|B)P(B)$

Naive Bayes assumption

Naive Bayes assumption:

$$P(x_1 \mid y, x_2) = P(x_1|y)\,.$$

Chain rule of probabilities:

$$P(A, B) = P(A|B)P(B)$$

We want to calculate $P(x|y) = P(x_1, \ldots, x_d|y)$.

Naive Bayes classifier

$$P(x \mid y) = P(x_1, \ldots, x_d | y) = \prod_{i=1}^{d} P(x_j \mid y)$$

Our model is parametrized (the stuff we need to remember after training) by

$$\phi_{i|y=1} = P(x_i = 1 \mid y = 1)\,,$$
$$\phi_{i|y=0} = P(x_i = 1 \mid y = 0)\,,$$
$$\phi_{y=1} = P(y = 1)\,.$$

By Bayes rule we get for a feature $x \in \mathcal{X}$

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1) P(y = 1)}{P(x)}$$

$$P(x) = P(x \mid y = 1) P(y = 1) + P(x \mid y = 1) P(y = 1)$$

Naive Bayes classifier: Training

$$\phi_{i|y=1} = P(x_i = 1 \mid y = 1)\,,$$
$$\phi_{i|y=0} = P(x_i = 1 \mid y = 0)\,,$$
$$\phi_{y=1} = P(y = 1)\,.$$

Indicator function

$$I(S) = \begin{cases} 1\,, & S \text{ is true} \\ 0\,, & S \text{ is false} \end{cases}\,.$$

Given a training set $\mathcal{T} = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$ we can calculate them by

$$\phi_{i|y=1} = \frac{\sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 1)}{\sum_{j=1}^{n} I(y^{(j)} = 1)}$$

$$\phi_{i|y=0} = \frac{\sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 0)}{\sum_{j=1}^{n} I(y^{(j)} = 0)}$$

$$\phi_{y=1} = \frac{1}{n} \sum_{j=1}^{n} I(y^{(j)} = 1)$$