Introduction to Machine Learning Course

Prof. Mingkui Tan

SCUT Machine Intelligence Laboratory (SMIL)

Contents

课程教学大纲

- 机器学习基础（3）
- Linear Regression and Gradient Descent （3）
  线性回归与梯度下降
- Linear Classification and Stochastic Gradient Descent （3）
  线性分类、支持向量机、随机梯度算法
- Logistic Regression and Ensemble Methods (Decision Tree, Adaboost) （3）
  逻辑回归与集成学习算法
- Overfitting, Underfitting, Regularization and Cross-Validation (3)
  过拟合、欠拟合、正则化与交叉验证
- Multiclass Classification and Cross-entropy Loss （3）
  多类分类和交叉熵损失函数

课程教学大纲

- ~~Clustering and Dimension Reduction (PCA, Feature Selection) （3）~~
  ~~聚类算法与维度约简~~
- ~~**Recommendation Systems （3） 推荐系统**~~
- Neural Networks and Deep Learning (Basics) （3）
  神经网络与深度学习
- Image Processing Basics and Convolutional Neural Networks （3）
  神经网络与深度学习
- 序列模型(RNN)、Transformer、Bert（3）
- ~~Markov Decision Process, Reinforcement Learning and AlphoGO （3）~~
  ~~马尔可夫决策过程、强化学习及AlphoGo~~

实验教学大纲

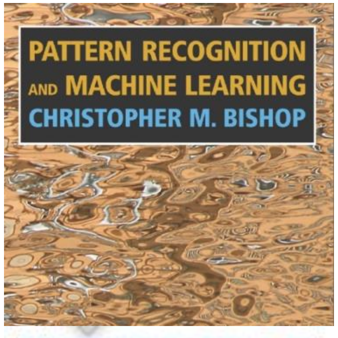
- **随堂实验**
  - Linear Regression and Gradient Descent (2)
    线性回归与梯度下降
  - Linear Classification with Stochastic Gradient Descent (2)
    线性分类、支持向量机、随机梯度算法
- **课程实验**
  - Classification with AdaBoost (4)
    科技论文阅读、写作;
    逻辑回归与集成学习算法
  - Face Detection and Recognition (4)
    人脸检测与识别基础
  - 基于Transformer的中英文翻译 (4)

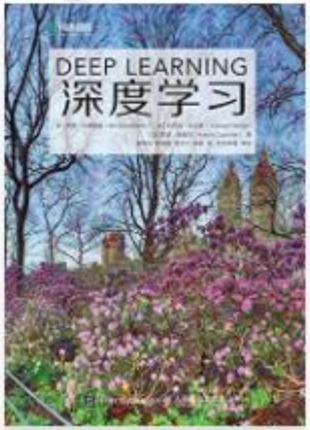
考核标准+参考书

■ 考核标准

考试（50%）+ **平时成绩（25%）** + **技术报告（25%）**

■ 参考书





- Pattern Recognition and Machine Learning **By Bishop**
- Understanding Machine Learning: From Theory to Algorithms By Shai Shalev-Shwartz and Shai Ben-David
- 深度学习 by Ian Goodfellow（伊恩·古德费洛）
- 《机器学习》By 周志华

联系方式

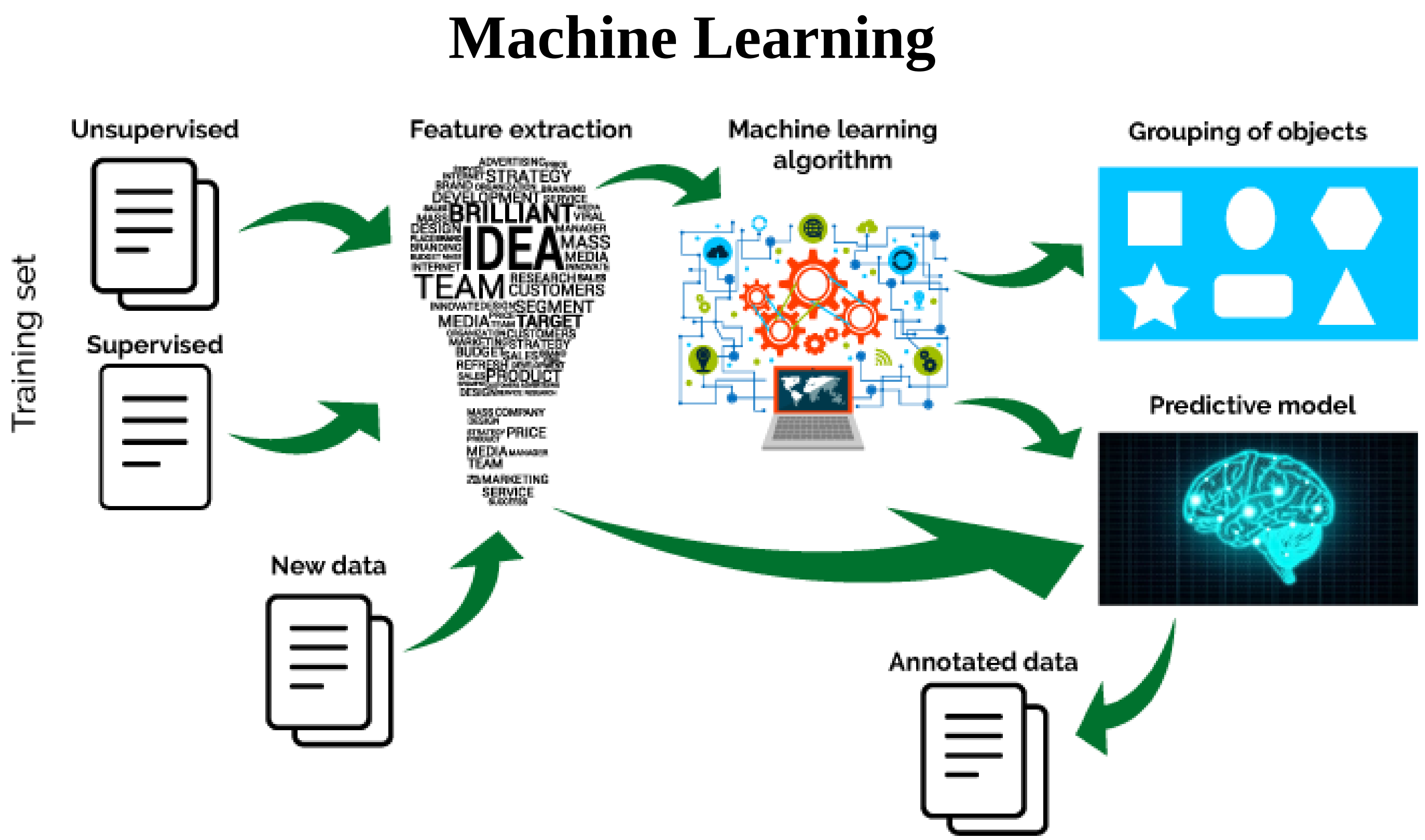
- **邮箱**

  mingkuitan@scut.edu.cn

- **QQ群**

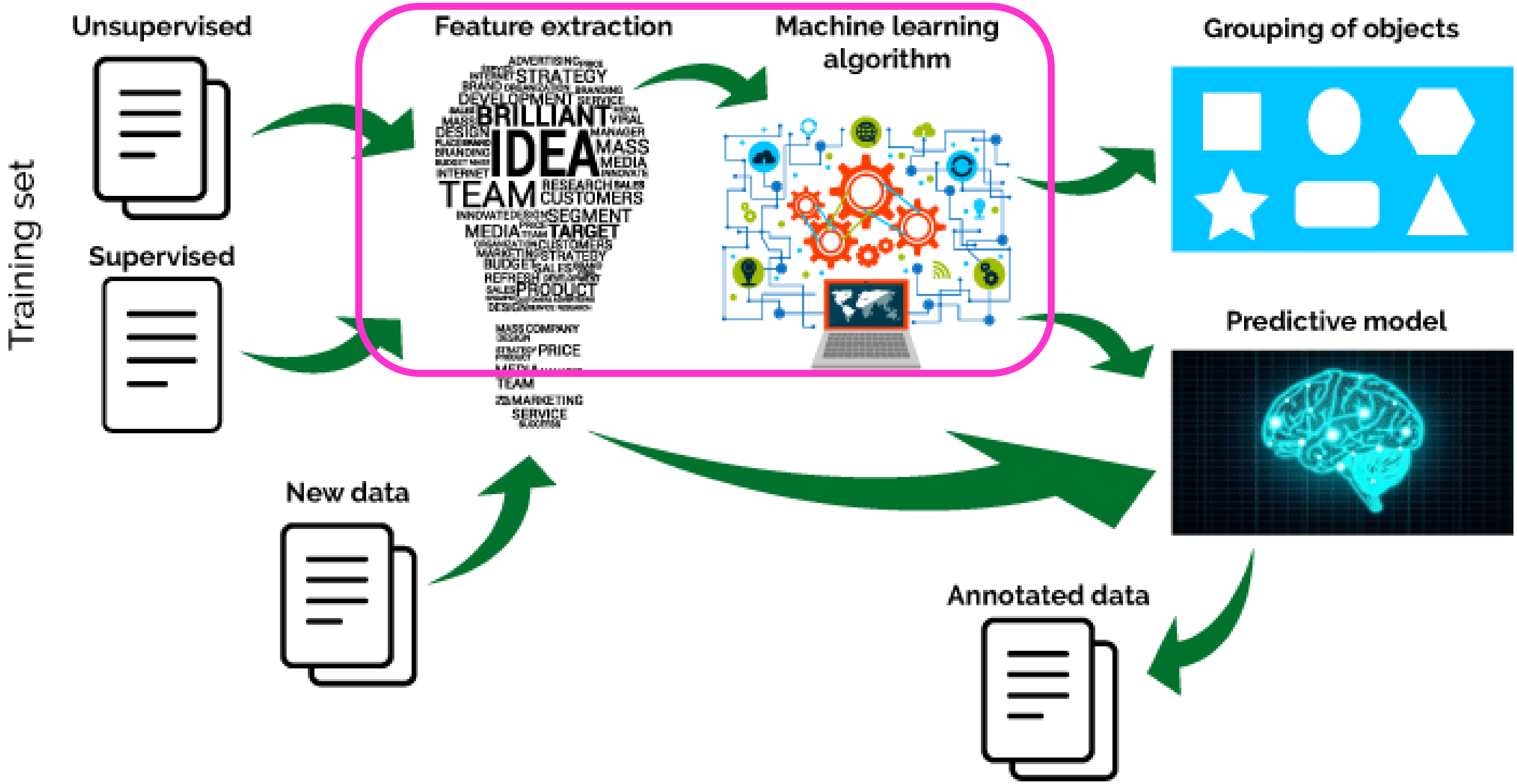
Contents

Big Picture: Machine Learning



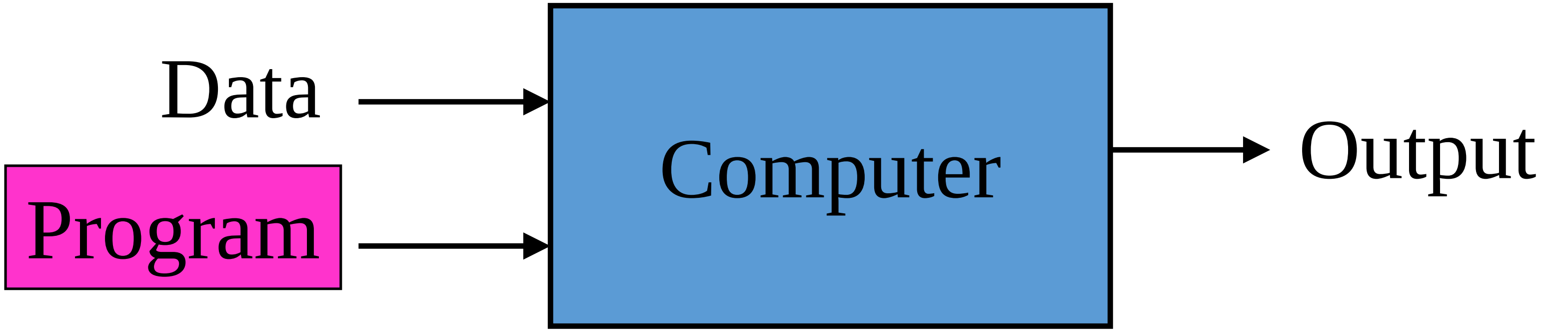
Big Picture: Machine Learning

Deep Learning

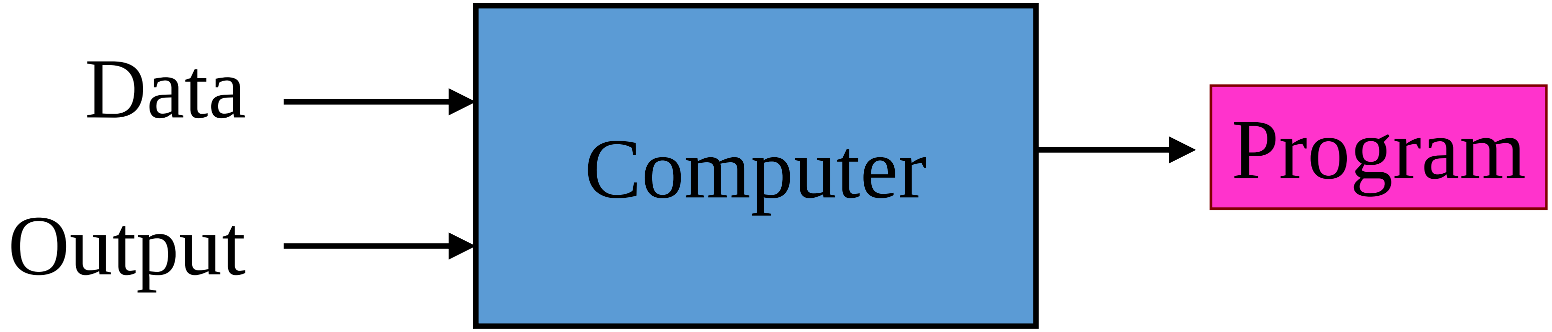


Traditional Programming and Machine Learning

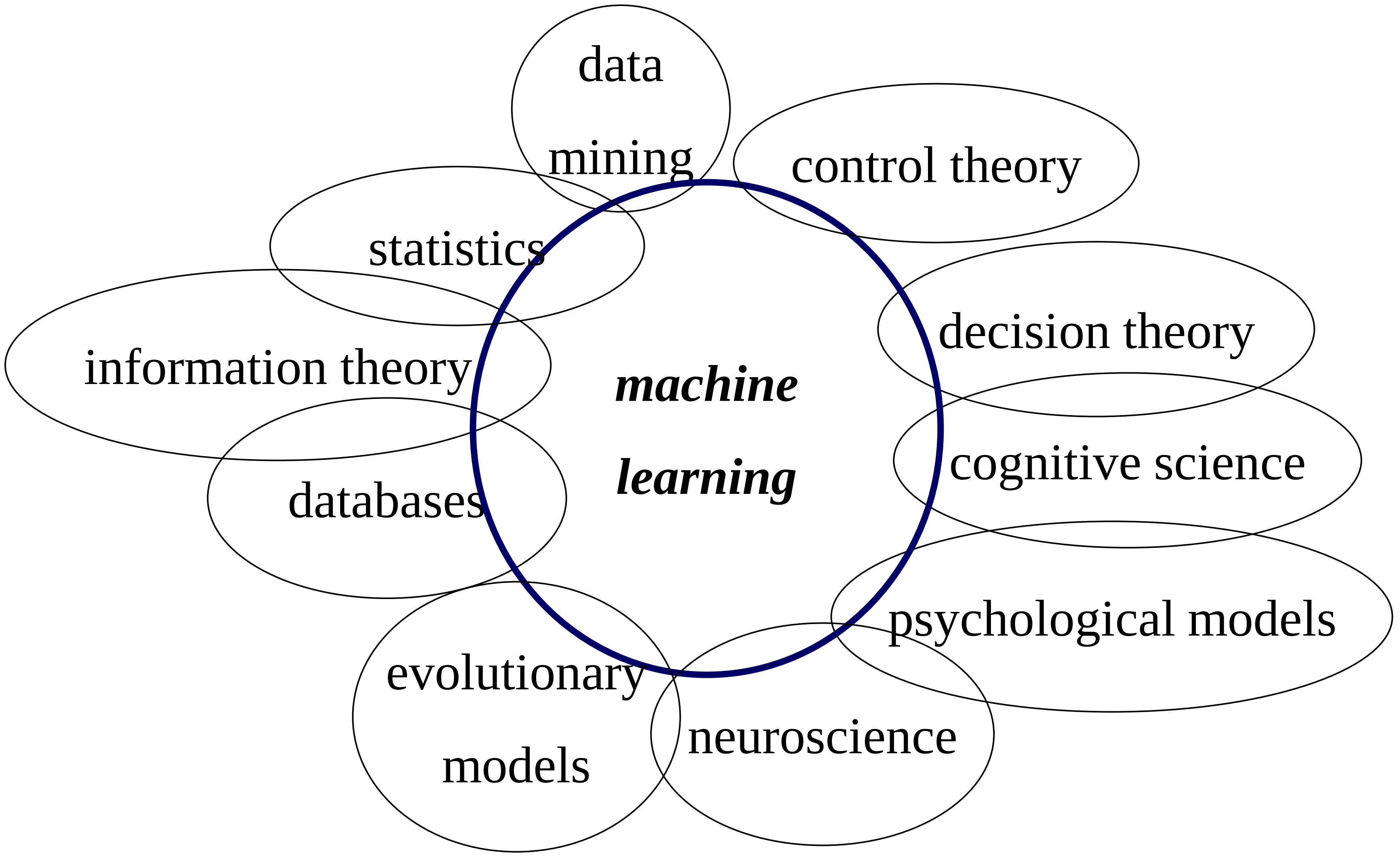
- **Traditional Programming**



- **Machine Learning**



Related Fields



Contents

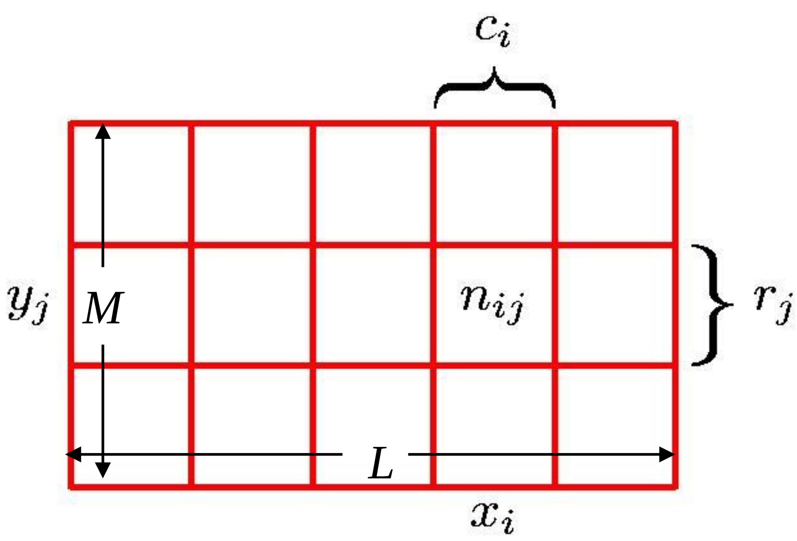
Probability Theory

- **Random Variables**

$$P(A) = \frac{1}{6}, A = 1, 2, \ldots, 6$$

- Random variables describe the outcome of a random experiment in terms of a (real) number
- A random experiment is an experiment that can (in principle) be repeated several times under the same conditions
- Discrete or continuous random variables
- Independent and identically distributed (iid) experiment vs non-iid experiment

Probability Theory



- **Marginal Probability**

$$P(X = x_i) = \frac{c_i}{L}$$

- **Joint Probability**

$$P(X = x_i, Y = y_i) = \frac{n_{ij}}{L \times M} = \frac{c_i \times r_j}{L \times M}$$

- **Conditional Probability**

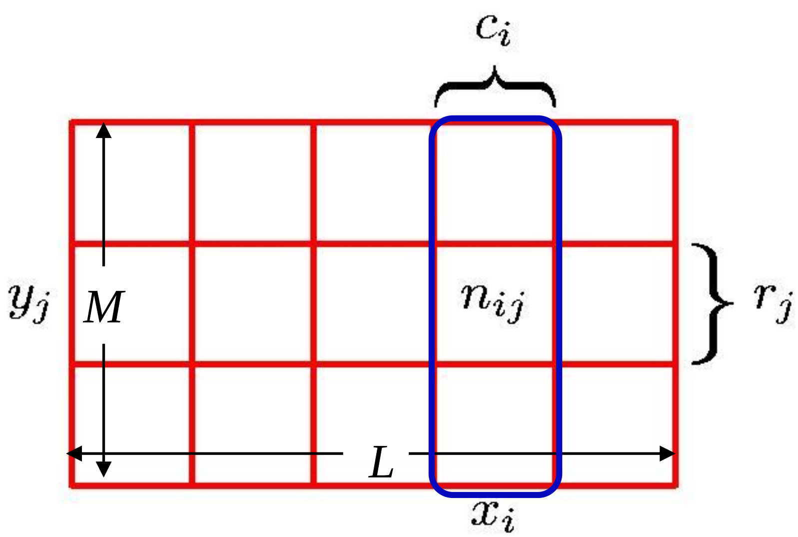
$$P\left(Y = y_j \mid X = x_i\right) = \frac{r_j}{M}$$

Probability Theory

- **Sum Rule**

$$P(X = x_i) = \frac{c_i}{L} = \frac{1}{L \times M} \sum_j n_{ij}$$

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



- **Product Rule**

$$P(X = x_i, Y = y_i) = \frac{n_{ij}}{L \times M} = \frac{r_j}{M} \cdot \frac{c_i}{L}$$

$$= P\left(Y = y_j \mid X = x_i\right) P(X = x_i)$$

Marginalization

Marginal Probability Joint Probability

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

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

Conditional Probability Marginal Probability

Y \ X	x_1	x_2	x_3	x_4	$p_y(Y)\downarrow$
y_1	$4/32$	$2/32$	$1/32$	$1/32$	$8/32$
y_2	$2/32$	$4/32$	$1/32$	$1/32$	$8/32$
y_3	$2/32$	$2/32$	$2/32$	$2/32$	$8/32$
y_4	$8/32$	0	0	0	$8/32$
$p_x(X) \rightarrow$	$16/32$	$8/32$	$4/32$	$4/32$	$32/32$

Margin

This concept is called "marginal" because it can be found by summing values in a table along rows or columns, and writing the sum in the **margins** of the table

Contents

Bayes' Theorem

The Rules of Probability

$$\text{Sum Rule: } P(X) = \sum_Y P(X, Y)$$

$$\text{Product Rule: } P(X, Y) = P(Y|X)P(X)$$

Bayes' Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \qquad P(X) = \sum_Y P(X|Y)P(Y)$$

Bayes' Theorem

posterior $\propto$ likelihood $\times$ prior

Posterior probability $P(Y|X)$: the likelihood of event Y occurring given that X is true, $P(Y|X)$ is a conditional probability

Posterior probability $P(X|Y)$: the likelihood of event X occurring given that Y is true, $P(X|Y)$ is a conditional probability

Prior probability $P(X)$ and $P(Y)$: the probabilities of observing X and Y independently of each other (the marginal probability)

Bayes' Theorem

$$P(\text{"taking a shower"}|\text{"wet"}) = P(\text{"wet"}|\text{"taking a shower"})\frac{P(\text{"taking a shower"})}{P(\text{"wet"})}$$

$$P(\text{reason}|\text{observation}) = P(\text{observation}|\text{reason})\frac{P(\text{reason})}{P(\text{observation})}$$

- Often useful in diagnosis situations,
  since $P(\text{observation}|\text{reason})$ might be easily determined
- Useful for reasoning
- Often delivers surprising results

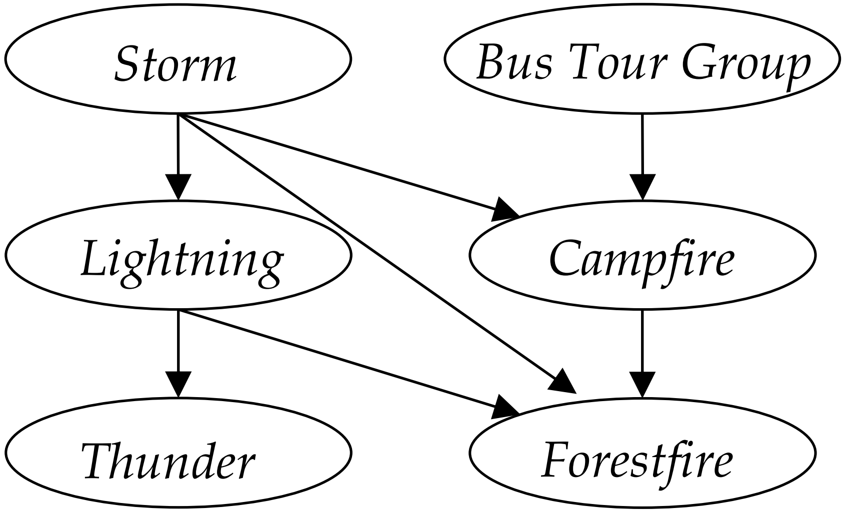
Bayes' Theorem in Bayesian Learning

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$: prior probability of hypothesis h
- $P(D)$: prior probability of training data D
- $P(h|D)$: posterior probability of h given D
- $P(D|h)$: posterior probability of D given h

Bayesian Net

- Network represents conditional independence assertions
- Each node conditionally independent of its non-descendants, given its immediate predecessors (e.g. Campfire and Lightning are independence conditioned on Storm)



conditional probability tables (CPT)

	$S \wedge B$	$S \wedge \neg B$	$\neg S \wedge B$	$\neg S \wedge \neg B$
C	0.4	0.1	0.8	0.2
$\neg C$	0.6	0.9	0.2	0.8

C: Campfire

S: Storm

B: Bus Tour Group

Example

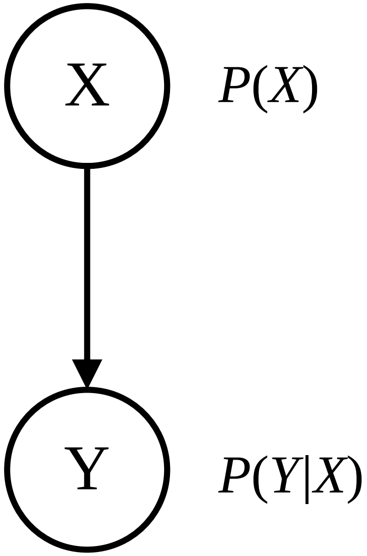
- **Random variables *X* and *Y***

  X: It is raining
  Y: The grass is wet

- ***X* affects *Y***

  Or, *Y* is a symptom of *X*

- **Draw two nodes and link them**
  - Define the CPT(conditional probability tables) for each node
    - $P(\mathrm{X})$ and $P(Y|X)$
  - Typical use: we observe Y and we want to query $P(X|Y)$
  - - Y is an evidence variable
    - X is a query variable



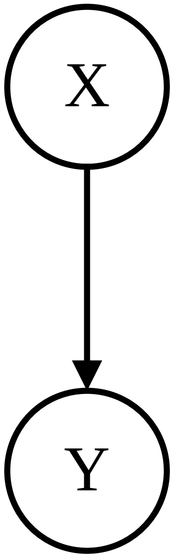
Example

- **What is $P(X|Y)$?**
  - Given that we know the CPTs of each node in the graph

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

$$= \frac{P(Y \mid X)P(X)}{\sum_X P(X, Y)}$$

$$= \frac{P(Y \mid X)P(X)}{\sum_X P(Y \mid X)P(X)}$$

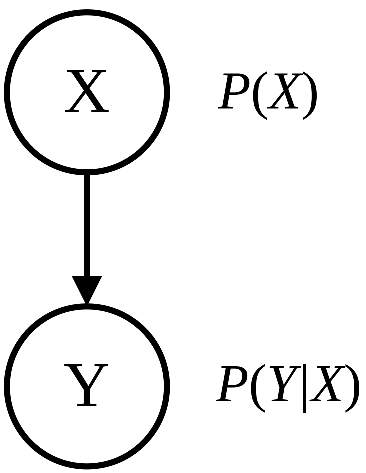


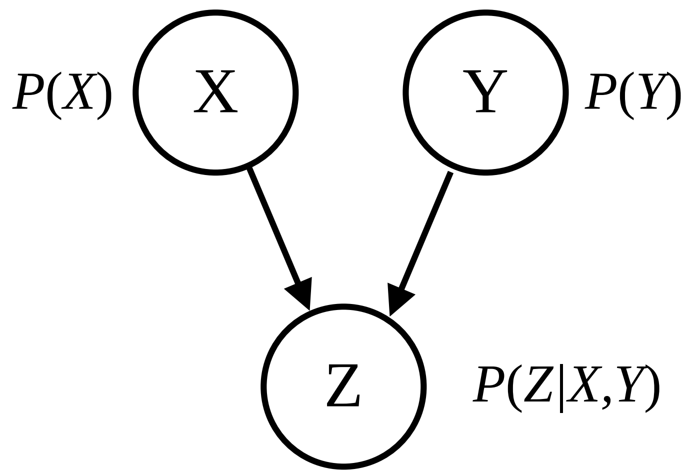
$P(X)$

$P(Y|X)$

Belief Nets Represent Joint Probability

- **The joint probability function can be calculated directly from the network**
- **It is the product of the CPTs of all the nodes**
- $\boldsymbol{P(var_1, \ldots, var_n) = \prod_i P(var_i | \text{Parents}(var_i))}$





$\boldsymbol{P(X,Y) = P(X)P(Y|X)}$

$\boldsymbol{P(X,Y,Z) = P(X)\,P(Y)\,P(Z|X,Y)}$

Probability Densities

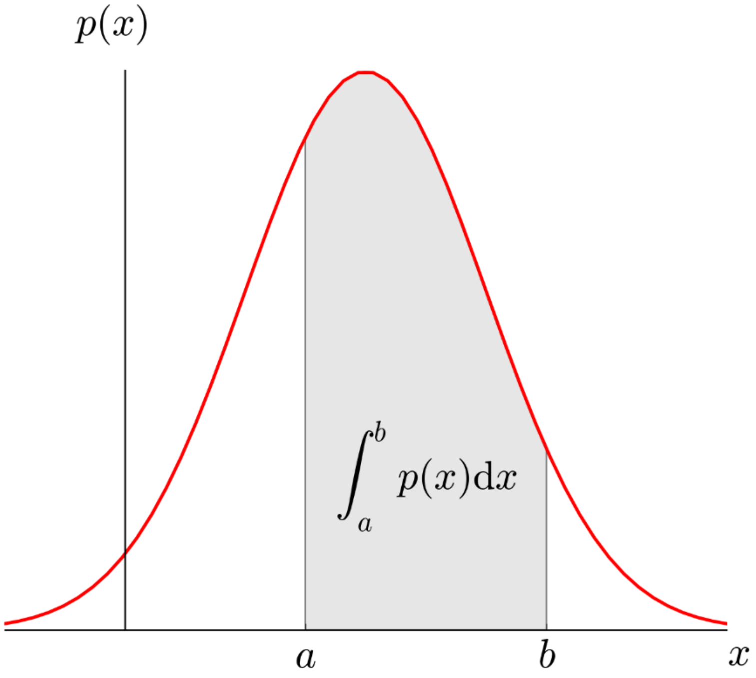
- **The probability density function p(x) has the following properties**

$$p(x) \geqslant 0$$

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

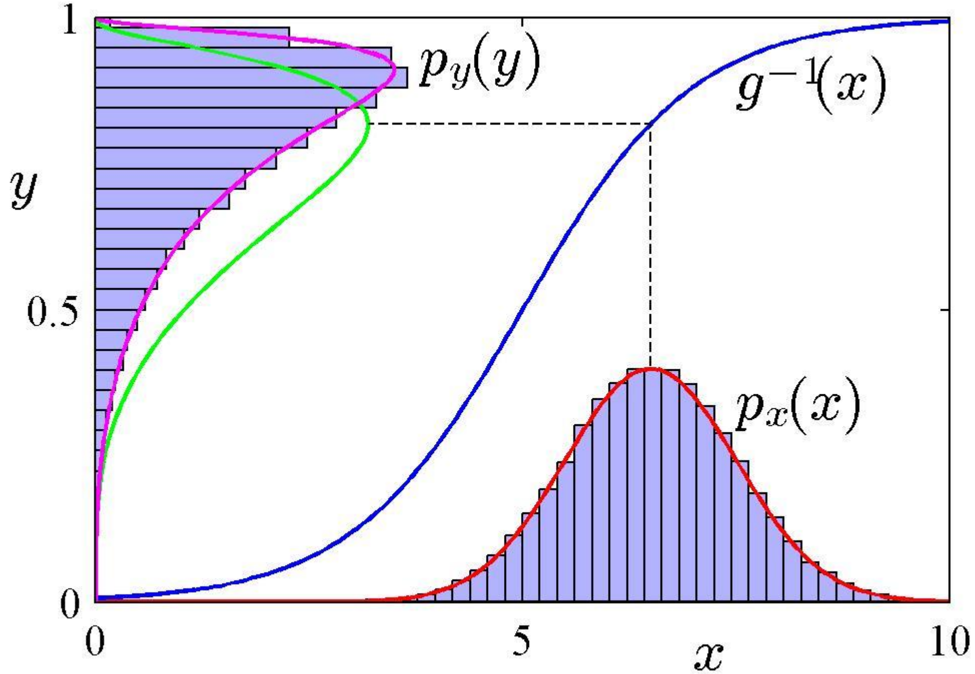
$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$



Transformed Densities

- x has a probability density $p_x(x)$
- $y = h(x)$ is some strictly monotonic continuous function
- Probability density $p_y(y)$ can be transformed from $p_x(x)$



$$y = h(x) = g^{-1}(x)$$

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\,|g'(y)| \end{aligned}$$

Maximum Likelihood Estimation

- **A density $\boldsymbol{f}$ usually contains parameters $\boldsymbol{\theta} \in \Omega$: $\boldsymbol{f(x|\theta)}$**
  Parameters $\boldsymbol{\theta}$: a scalar or a vector

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **Question: How to estimate $\boldsymbol{\theta}$ given data $\mathcal{D} = \{\boldsymbol{x_i}\}$?**
- **Likelihood function of $\boldsymbol{\theta}$ given $\boldsymbol{x}$:**

$$L(\theta|x) = P(X = x|\theta)$$

- **Likelihood function of $\boldsymbol{\theta}$ given $\mathcal{D} = \{\boldsymbol{x_i}\}$:**

$$L_{\mathcal{D}}(\theta) = P(\mathcal{D}|\theta) = \prod_{m} P(x_i|\theta)$$

Maximum Likelihood Estimation

- **Likelihood function of $\boldsymbol{\theta}$ given $\boldsymbol{\mathcal{D}} = \{\boldsymbol{x_i}\}$ (iid $\boldsymbol{x_i}$)**

$$L_{\mathcal{D}}(\theta) = P(\mathcal{D}|\theta) = \prod_i P(x_i|\theta)$$

- **Estimate $\boldsymbol{\theta}$ by**

$$\theta_* = \underset{\theta}{\operatorname{argmax}} \left(\prod_i P(x_i|\theta)\right)$$

- **In practice, often use log likelihood function**

$$\theta_* = \underset{\theta}{\operatorname{argmax}} \log\left(\prod_i P(x_i|\theta)\right)$$

- **Then, we have**

$$\theta_* = \underset{\theta}{\operatorname{argmax}} \left(\sum \log\big(P(x_i|\theta)\big)\right)$$

Maximum a Posteriori Estimation

- **Replace the likelihood in the MLE formula with the posterior, and we get:**

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} P(X|\theta)P(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \log P(X|\theta) + \log P(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \log \prod_i P(x_i|\theta) + \log P(\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_i \log P(x_i|\theta) + \log P(\theta)$$

MLE vs MAP

- **If we use uniform prior in MAP estimation, $P(\theta)$ is a const, so we have:**

$$\theta_{MAP} = \text{argmax} \sum_i \log P(x_i|\theta) + \log P(\theta)$$
$$= \text{argmax} \sum_i \log P(x_i|\theta) + const$$
$$= \text{argmax} \sum_i \log P(x_i|\theta) = \theta_{MLE}$$

- **MLE is a special case of MAP, where the prior is uniform**

Contents

Probability and Information Theory

- **Information measure of an event *A***

$$I(A) = -\log_b P(A)$$

***I*(*A*):** self-information or information content, random variable

***P*(*A*):** probability of the event happening

***b*:** base, usually b=2

base 2 = bits   base 3 = trits
base 10 = Hartleys base e = nats

Information and Probability

Examples

The Chinese football team lost：

$$P(A)=1 \qquad\qquad I(A) = -\log_2 P(A) = 0$$

The Chinese table tennis team lost：

$$P(A)=0 \qquad\qquad I(A) = -\log_2 P(A) = +\infty$$

Probability $P(A)$: The degree of uncertainty of an event

Self-information $I(A)$: The elimination of uncertainty

Entropy

- **Entropy is simply the average (expected) amount of the information from the event**

$$H(A) = -E[\log_2 P(A)] = -\sum_A P(A)\log_2 P(A)$$

H*(*A*) is maximized when $P(A) = \frac{1}{n}$ for all *A

- **Joint Entropy**

$$H(A,B) = -E[\log_2 P(A,B)] = -\sum_{A,B} P(A,B)\log_2 P(A,B)$$

- **Conditional entropy of *A* given *B***

$$H(A|B) = -E[\log_2 P(A|B)] = -\sum_{A,B} P(A,B)\log_2 P(A|B)$$

Thank You