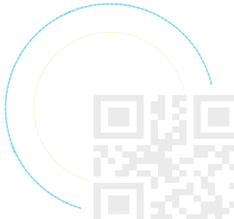
Mathematical Foundations of NLP

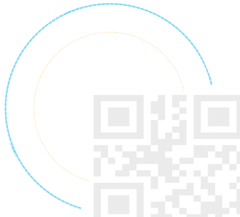
自然语言处理--数学基础

玖强

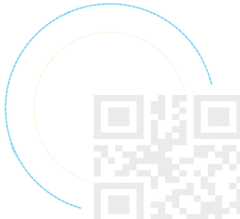


OUTLINE

- 概率和信息论
- 分类与回归模型
- 监督学习、半监督学习和非监督学习

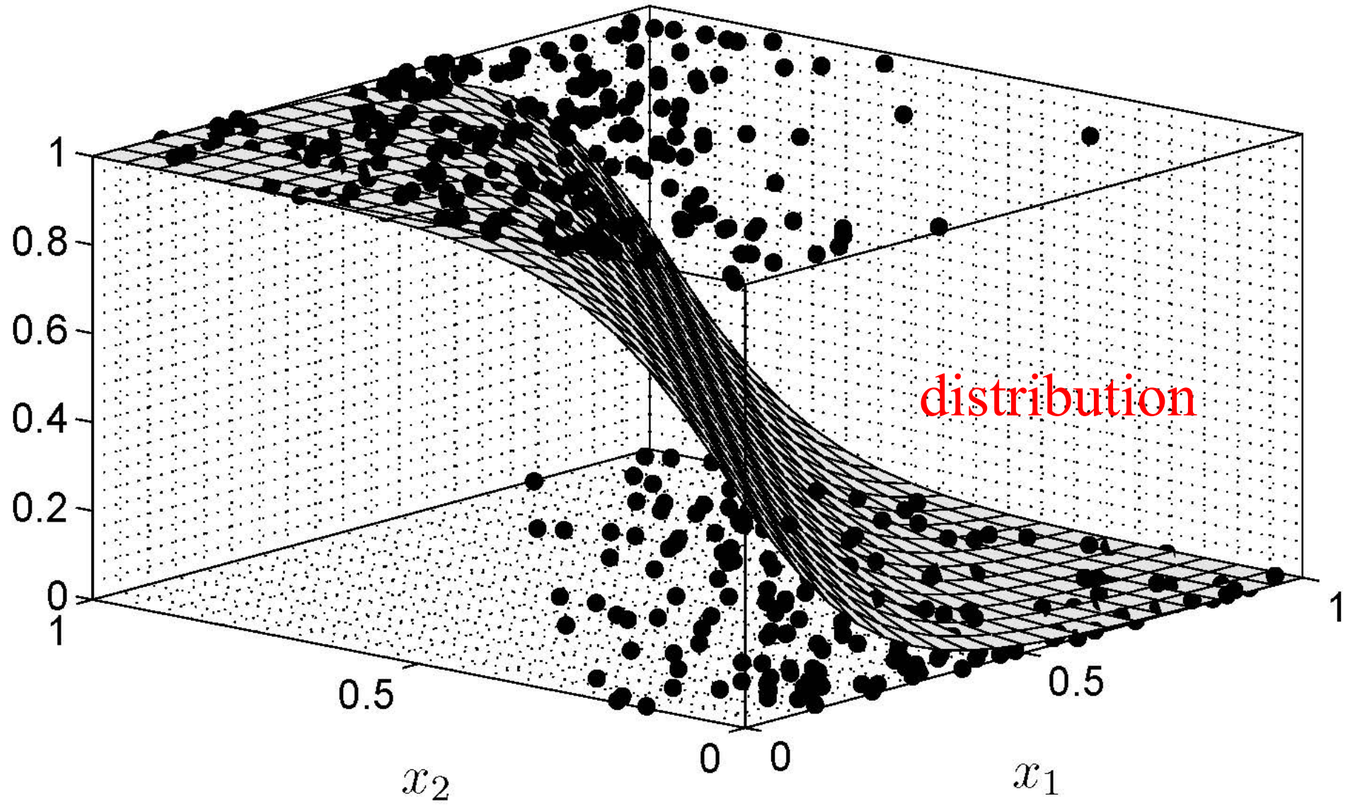


Probability and Information Theory for Language Modeling



Motivations

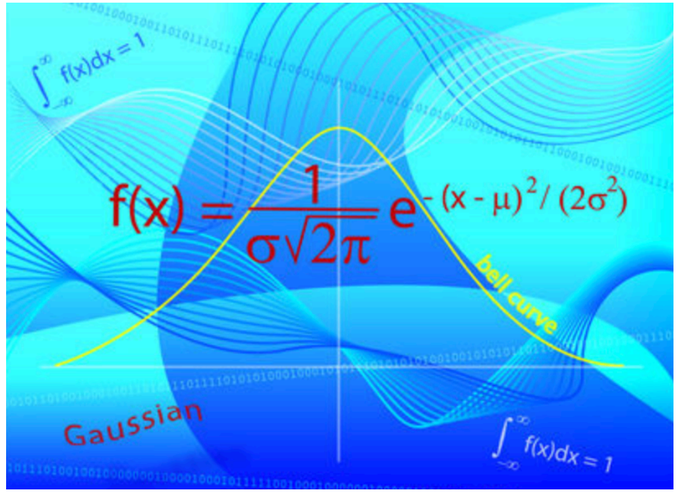
- Statistical NLP aims to do statistical inference for the field of NLP
- Statistical inference consists of taking some data (*generated in accordance with some unknown probability distribution*) and then making some inference about this distribution.



Motivations (Cont/条件)

- An example of statistical inference is the task of language modeling
  - 例如，给定previous words，如何预测下一个单词？
  - 我们需要一个语言模型！

- 概率论是个好东西！
  - Probability theory helps us finding such model



Probability Theory

- 一件事情可能发生的可能性
- Sample space/采样空间 Ω is listing of all possible outcomes of an experiment
- 事件A是Ω的子集
- Probability function (or distribution) /条件概率

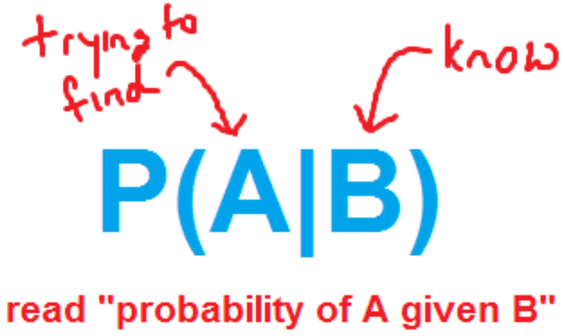
$$P: \Omega \rightarrow [0,1]$$

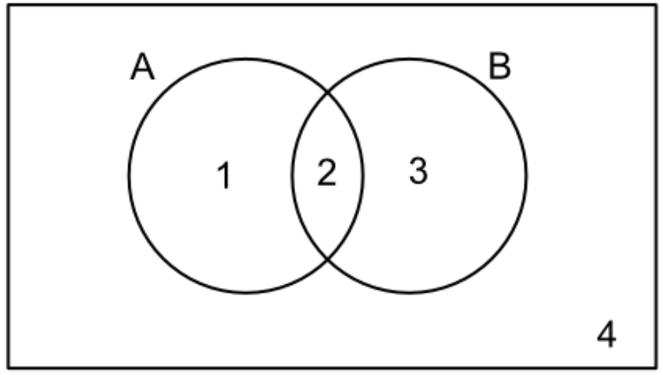
- Prior probability/先验概率
  - the probability before we consider any additional knowledge

$$P(A)$$

条件概率

- Sometimes we have partial knowledge about the outcome of an experiment
  - Conditional (or Posterior) Probability
- Suppose we know that event B is true
- The probability that A is true given the knowledge about B is expressed by : P(A|B)



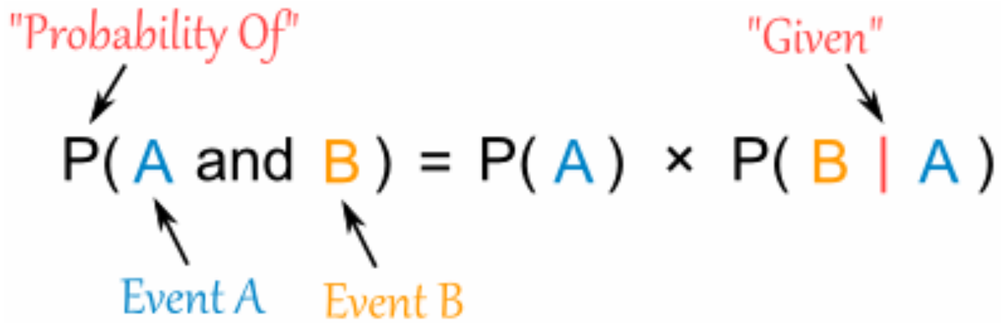


P (A | B) is A *given* B

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)} = \frac{2}{2+1} = \frac{2}{3}$$

Conditional probability (cont)

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

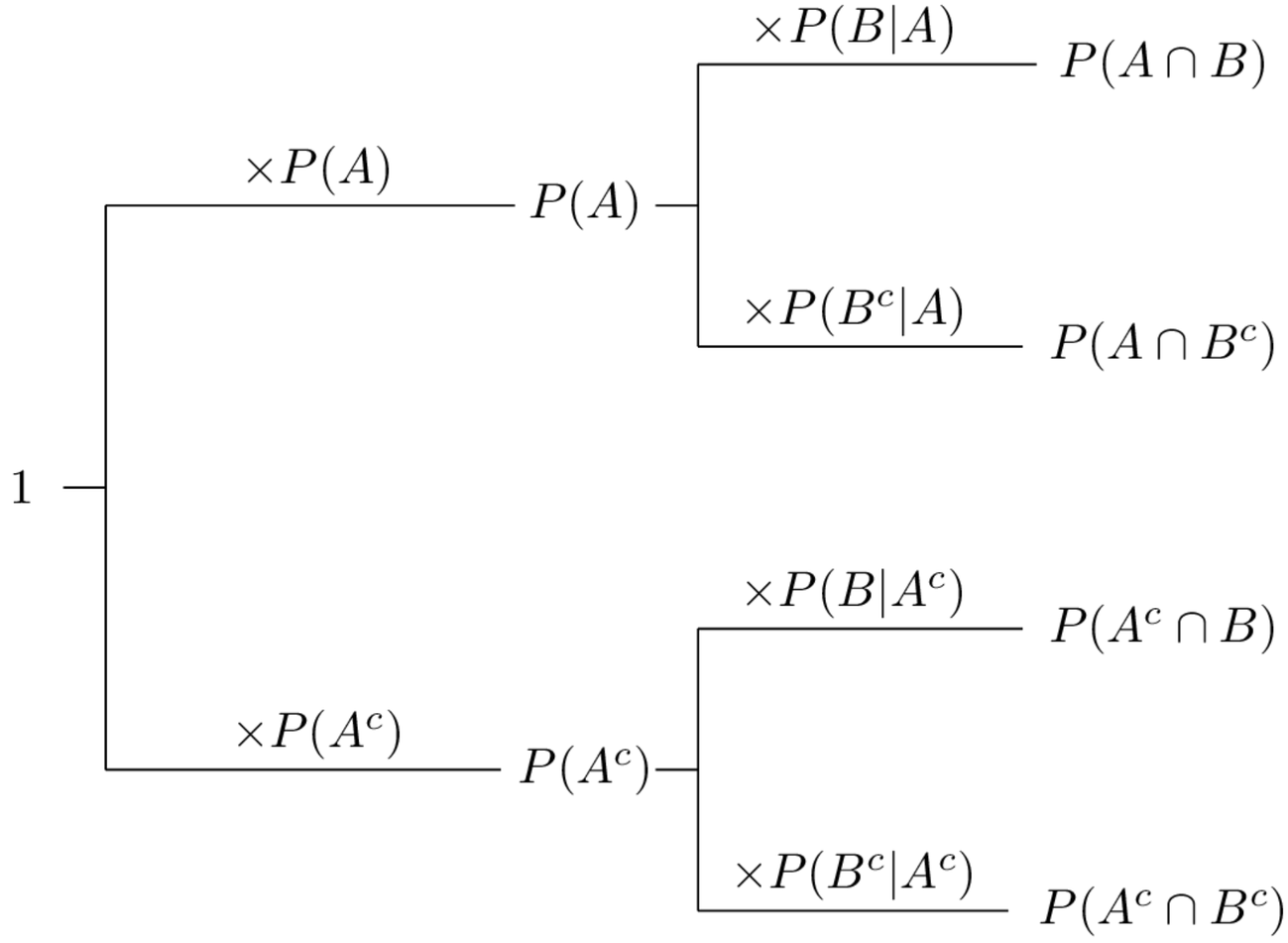


- Joint probability of A and B.

Chain Rule/链式法则

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A \cap B \cap C \cap D...) = P(A)P(B|A)P(C|A \cap B)P(D|A \cap B \cap C) ...$$



(Conditional) independence

- Two events A e B are *independent* of each other if

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

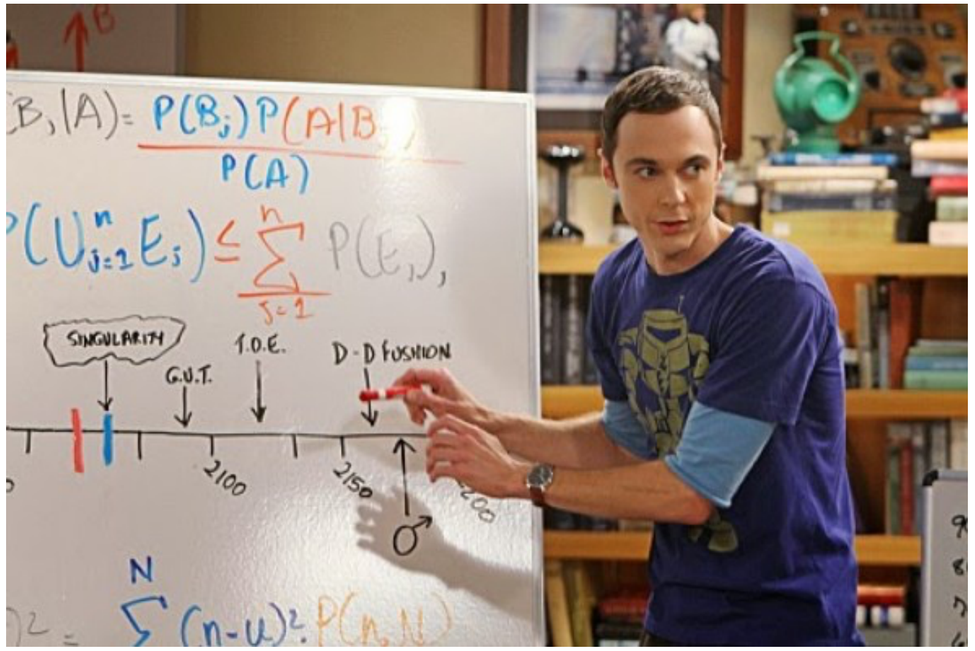
- Two events A and B are *conditionally independent of each other* given C if

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

Bayes' Theorem/贝叶斯理论

- 贝叶斯理论可以让我们交换事件之间的依赖顺序
- We saw that $P(A|B) = P(A \cap B)/P(B)$
- Bayes' Theorem:

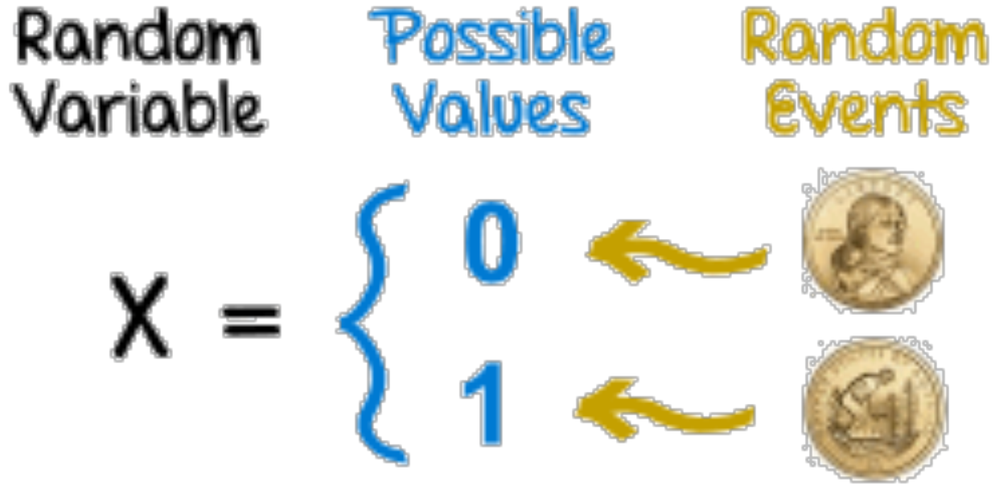
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Random Variables

- So far, event space that differs with every problem we look at
- Random variables (RV) X allow us to talk about the probabilities of values (numerical, categorical etc.) that are related to the event space :

$$X: \Omega \rightarrow \Re$$
$$X: \Omega \rightarrow \Im$$



Expectation / 期望

$$p(x) = p(X = x) = p(A_x)$$
$$A_x = \{\omega \in \Omega; X(\omega) = x\}$$

$$\sum_x p(x) = 1, \qquad 0 \leq p(x) \leq 1$$

- The Expectation of a RV is

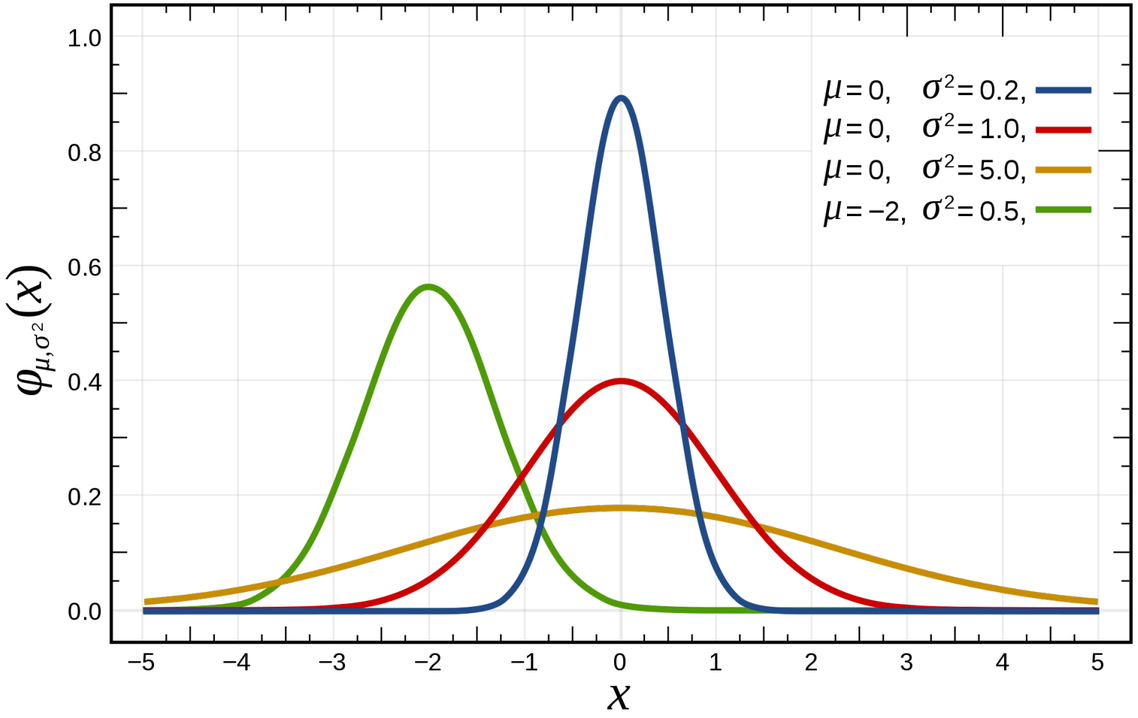
$$E(x) = \sum_x x p(x) = \mu$$

Variance / 方差

- The variance of a RV is a measure of the deviation of values of the RV about its expectation

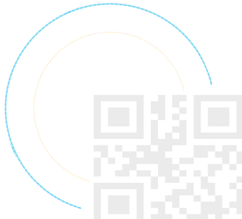
$$\mathrm{Var}(X)=\mathrm{E}((X-E(X))^2)=\mathrm{E}(X^2)\text{-}E^2(X)=\sigma^2$$

- σ is called the standard deviation



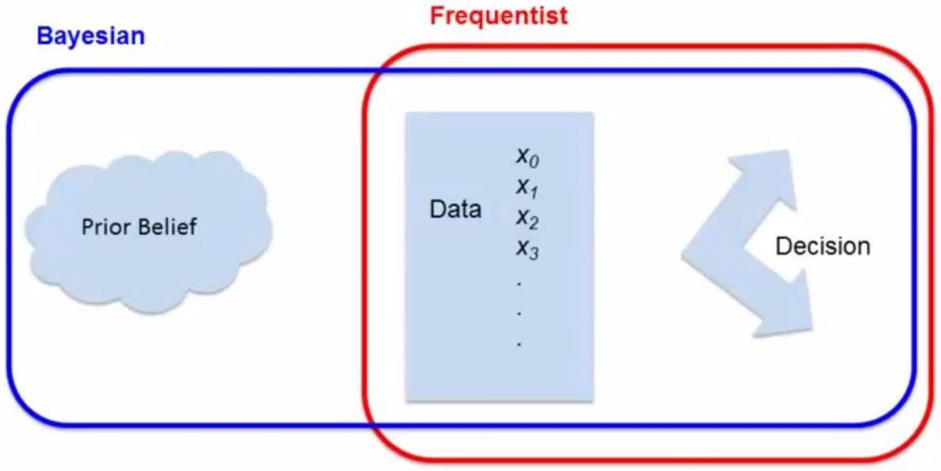
Back to the Language Model

- In general, for language events, P is unknown
- We need to estimate P, (or model M of the language)
- We'll do this by looking at evidence about what P must be based on a sample of data



Estimation of P

- Frequentist statistics / 频率学派
- Bayesian statistics / 贝叶斯学派
- 贝叶斯概率论为人的知识（knowledge）建模来定义「概率」这个概念。频率学派试图描述的是「事物本体」，而贝叶斯学派试图描述的是观察者知识状态在新的观测发生后如何更新。

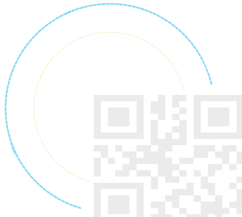


Frequentist Statistics

- Relative frequency: proportion of times an outcome u occurs

$$f_u = \frac{\mathrm{C(u)}}{\mathrm{N}}$$

- C(u) is the number of times u occurs in N trials
- For N→ ∞ the relative frequency tends to stabilize around some number: probability estimates
- Difficult to estimate if the number of different values u is large



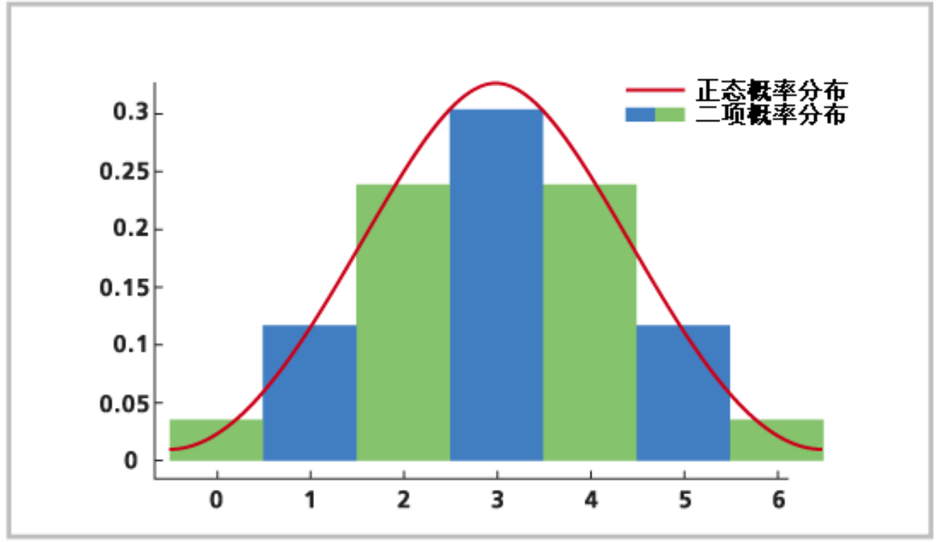
Binomial Distribution/ 二项分布 (Parametric)

- Series of trials with only two outcomes, each trial being independent from all the others
- 统计学家们总结出了计算概率的一般公式:

$$b(r;n,p) = \binom{n}{r} p^r (1-p^r)$$

符合以下4个特点的就是二项分布：

1. 做某件事的次数是固定的；
2. 每一件事情都有两个可能的结果（成功，或失败）
3. 每一次成功的概率都是相等的
4. 你最感兴趣的是成功x次的概率是多少

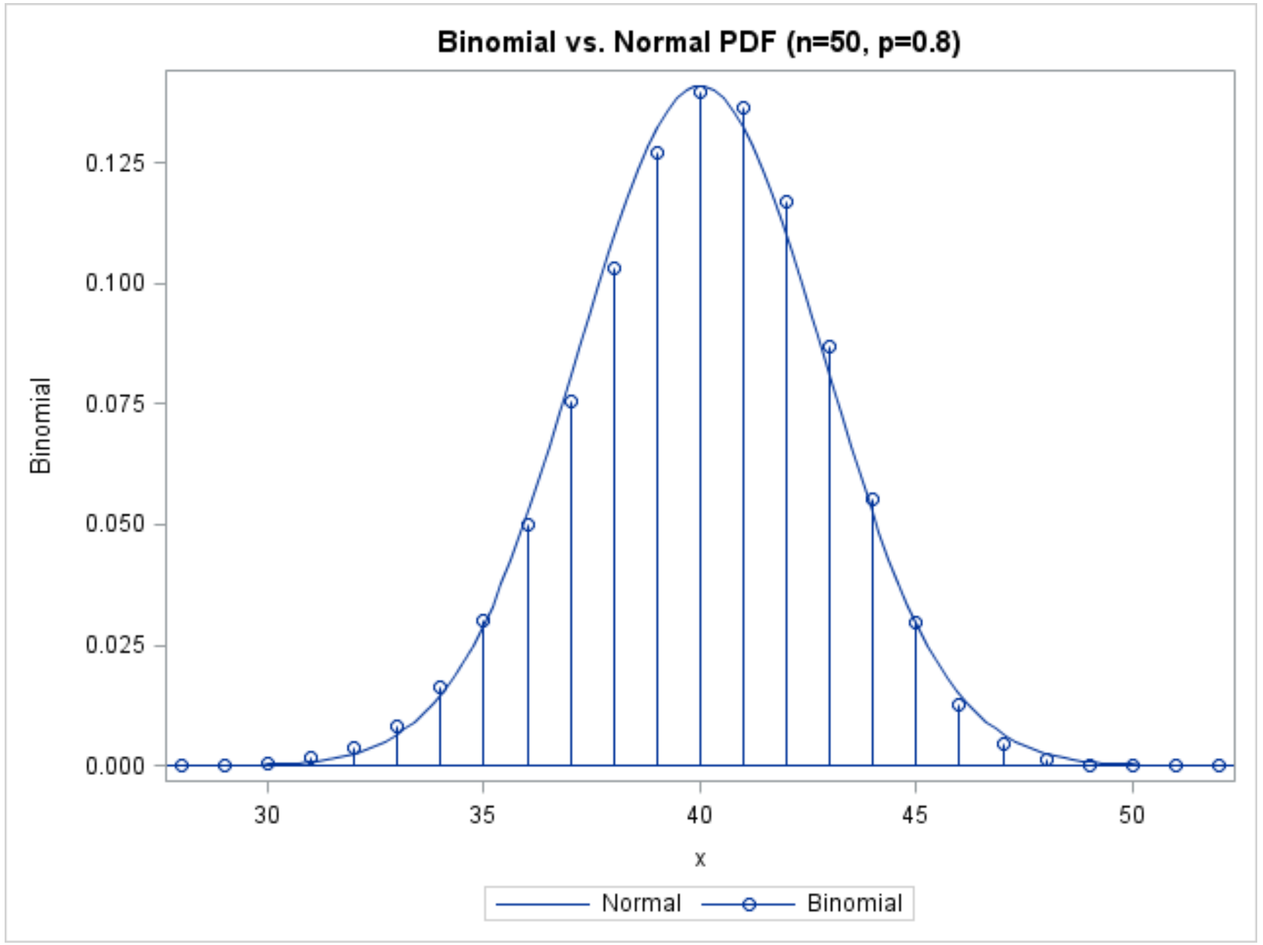


二项分布

Normal (Gaussian) Distribution (Parametric)

- Continuous
- Two parameters: mean μ and standard deviation σ

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

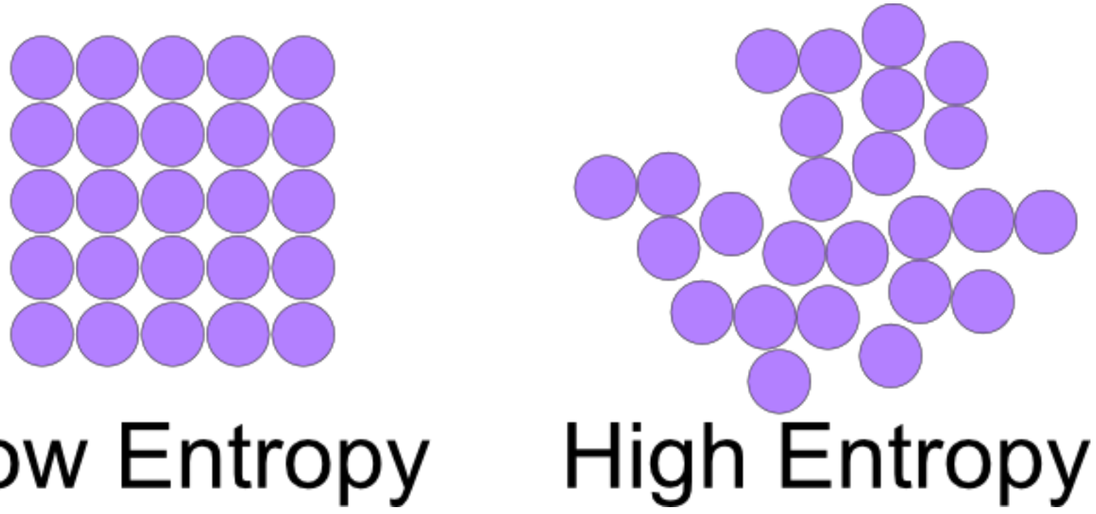


Entropy

- X: discrete RV, p(X)
- Entropy (or self-information)

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Entropy measures the amount of information in a RV; it's the average length of the message needed to transmit an outcome of that variable using the optimal code



Joint Entropy

- The joint entropy of 2 RV X,Y is the amount of the information needed on average to specify both their values

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

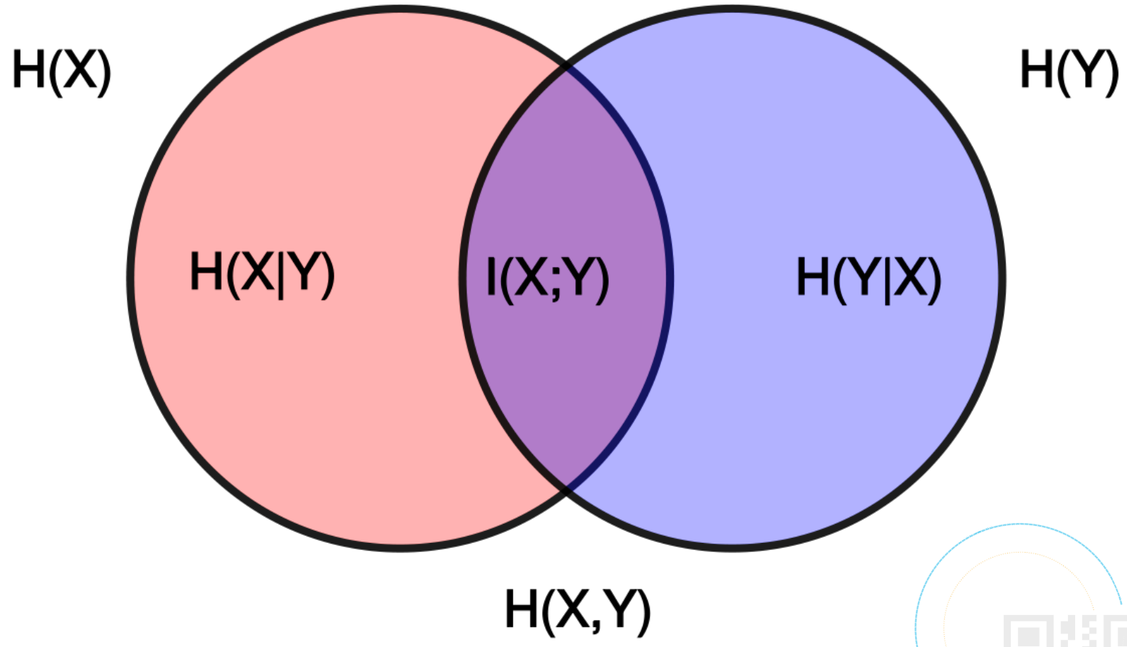
Chain rule

The joint entropy can be written as the sum

$$H(X, Y) = H(X) + H(Y|X)$$

Proof:

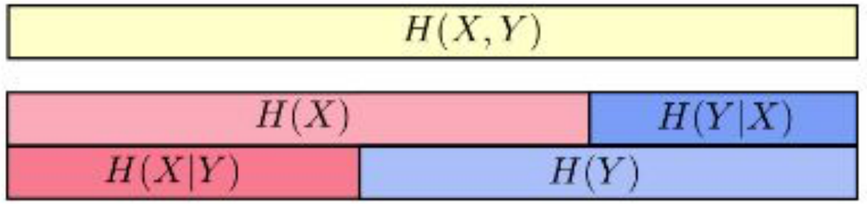
$$\begin{aligned} H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$



Conditional Entropy

- The conditional entropy of a RV Y given another X, expresses how much extra information one still needs to supply on average to communicate Y given that the other party knows X

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in X} p(x) H(Y|X = x) \\
&= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x) \\
&= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x) = -E(\log p(Y|X))
\end{aligned}
$$

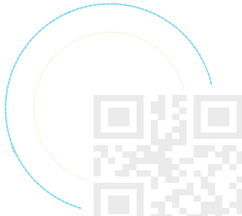


Mutual Information

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

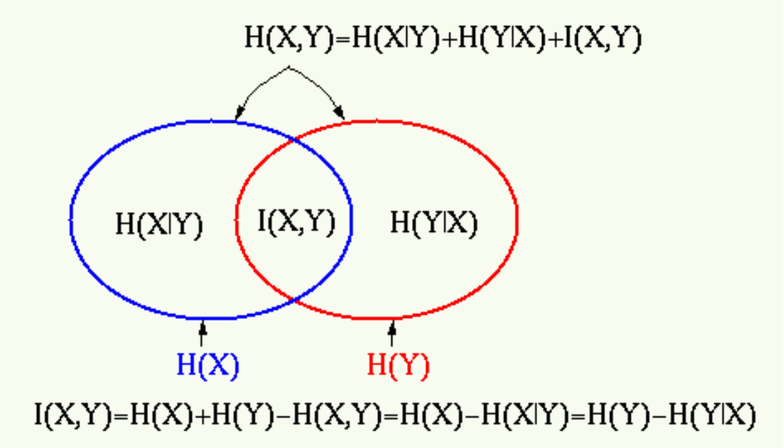
$$H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X,Y)$$

- I(X,Y) is the mutual information between X and Y. It is the reduction of uncertainty of one RV due to knowing about the other, or the amount of information one RV contains about the other



Mutual Information (cont)

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$



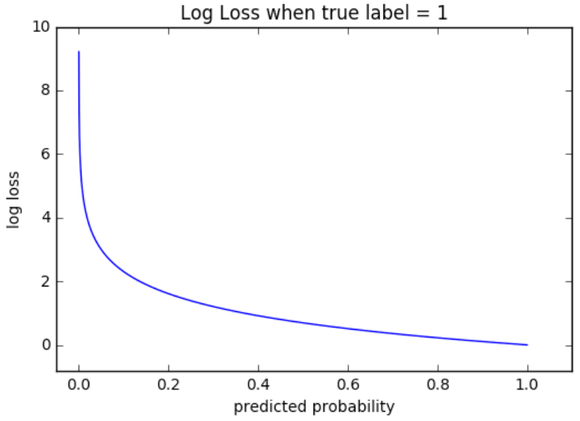
- I is 0 only when X,Y are independent: H(X|Y)=H(X)
- H(X)=H(X)-H(X|X)=I(X,X) Entropy is the self-information
- Maybe written as

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Entropy and Linguistics

- Entropy is measure of uncertainty. The more we know about something the lower the entropy.
- If a language model captures more of the structure of the language, then the entropy should be lower.
- We can use entropy as a measure of the quality of our models

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$



- H: entropy of language; we don't know p(X); so..?
- Suppose our model of the language is q(X)
- How good estimate of p(X) is q(X)?

$$\hat{\mathbf{y}} \qquad \mathbf{y}$$

$$\begin{bmatrix} 0.1 \\ 0.5 \\ 0.4 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

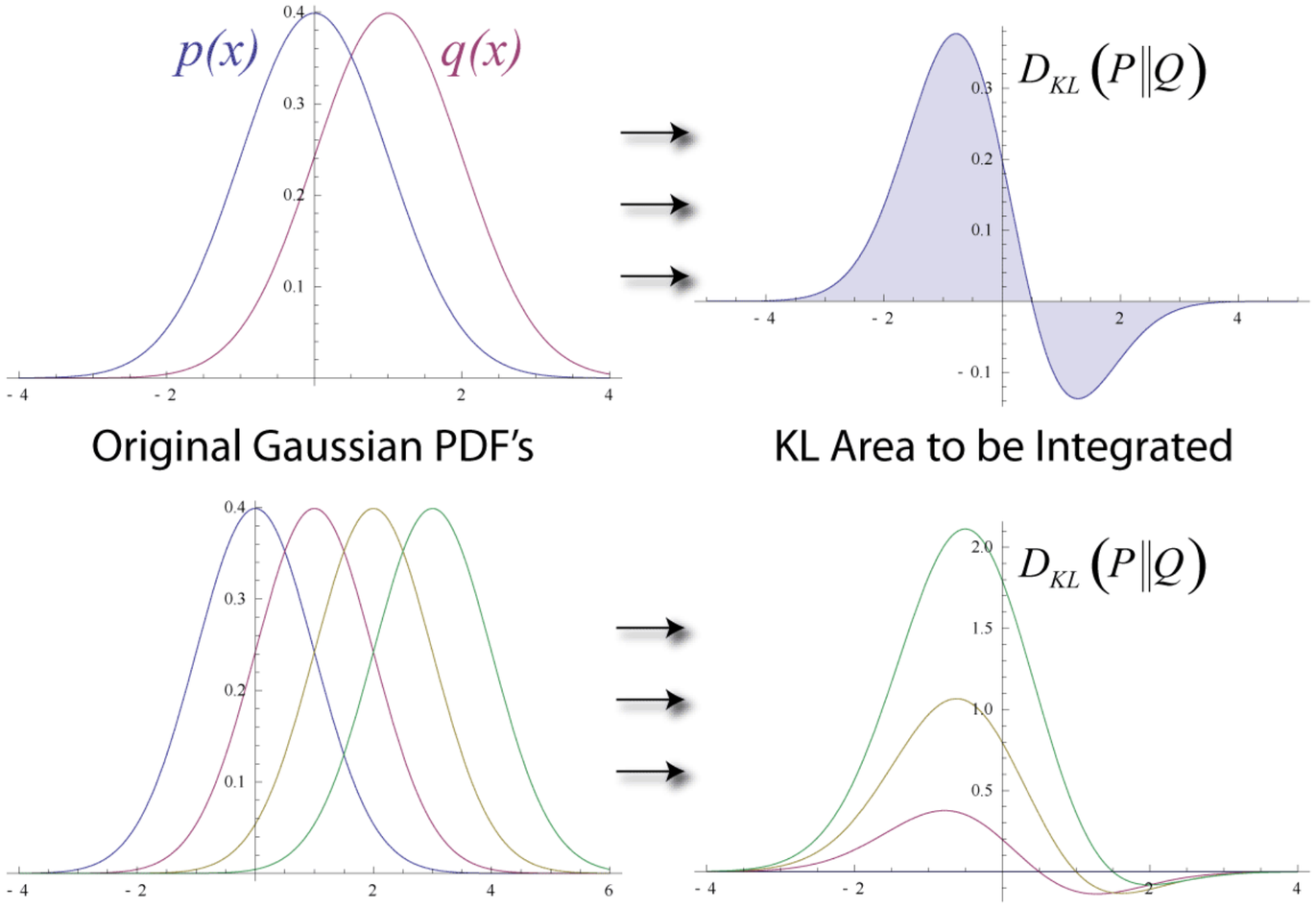
$$D(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_j y_j \ln \hat{y}_j$$

Entropy and Linguistics Kullback-Leibler Divergence

- Relative entropy or KL (Kullback-Leibler) divergence applies to two distributions p and q

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

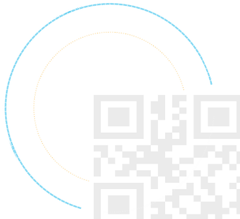
$$= E_p(\log \frac{p(X)}{q(X)})$$



VAE (KLD) → GAN (JKD)

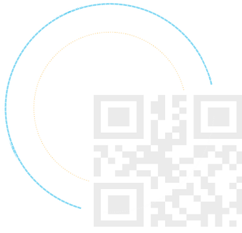
Unsupervised, Supervised and Semi-supervised Learning
监督学习、半监督学习和非监督学习

Classification and Regression
分类与回归模型



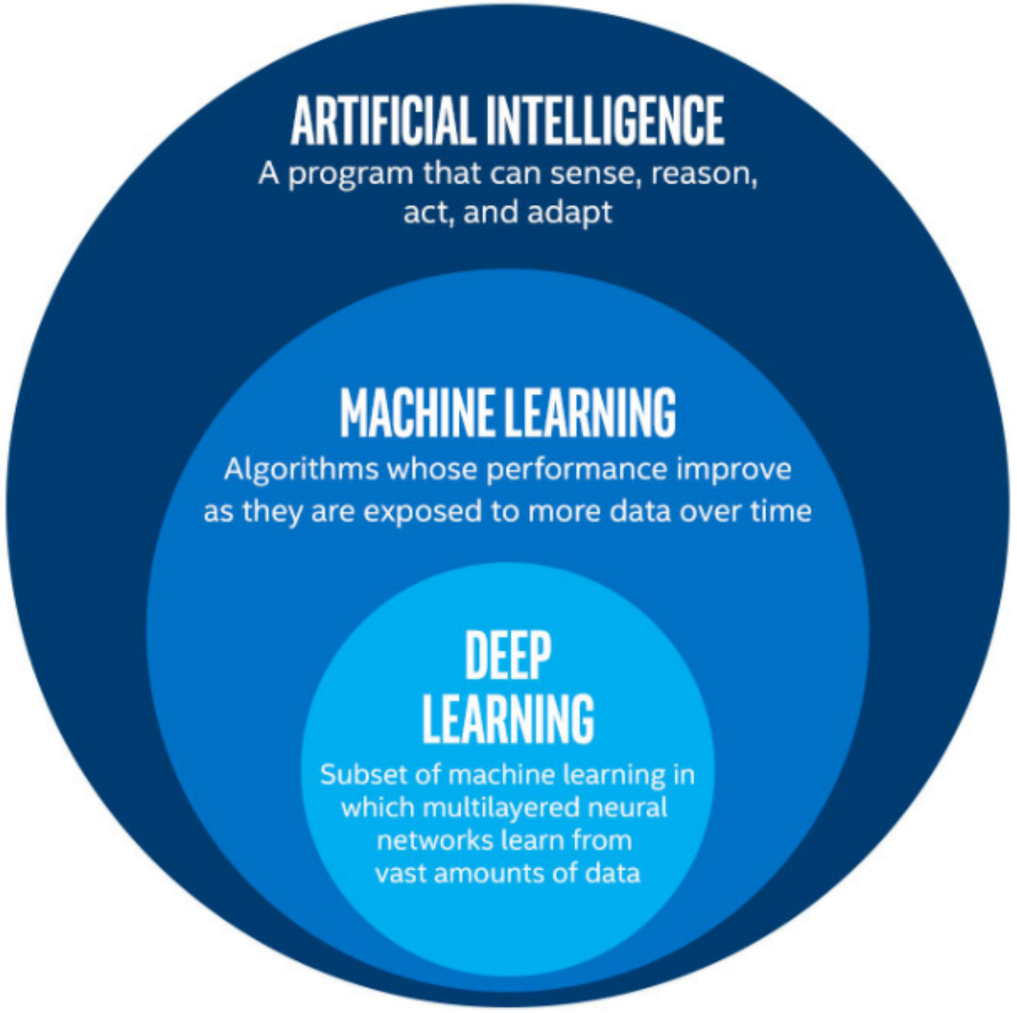
Why "Learn"?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to "learn" to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

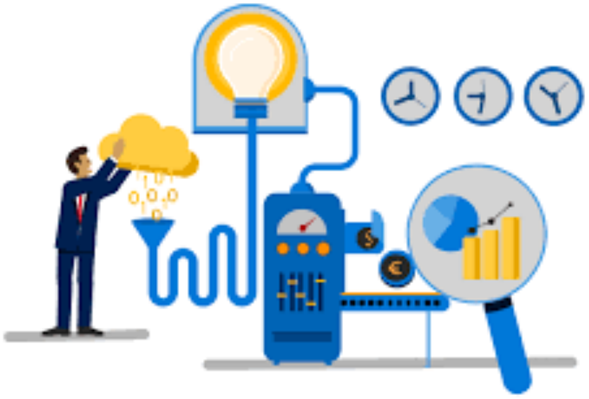


What We Talk About When We Talk About "Learning"

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Build a model that is a good and useful approximation to the data.



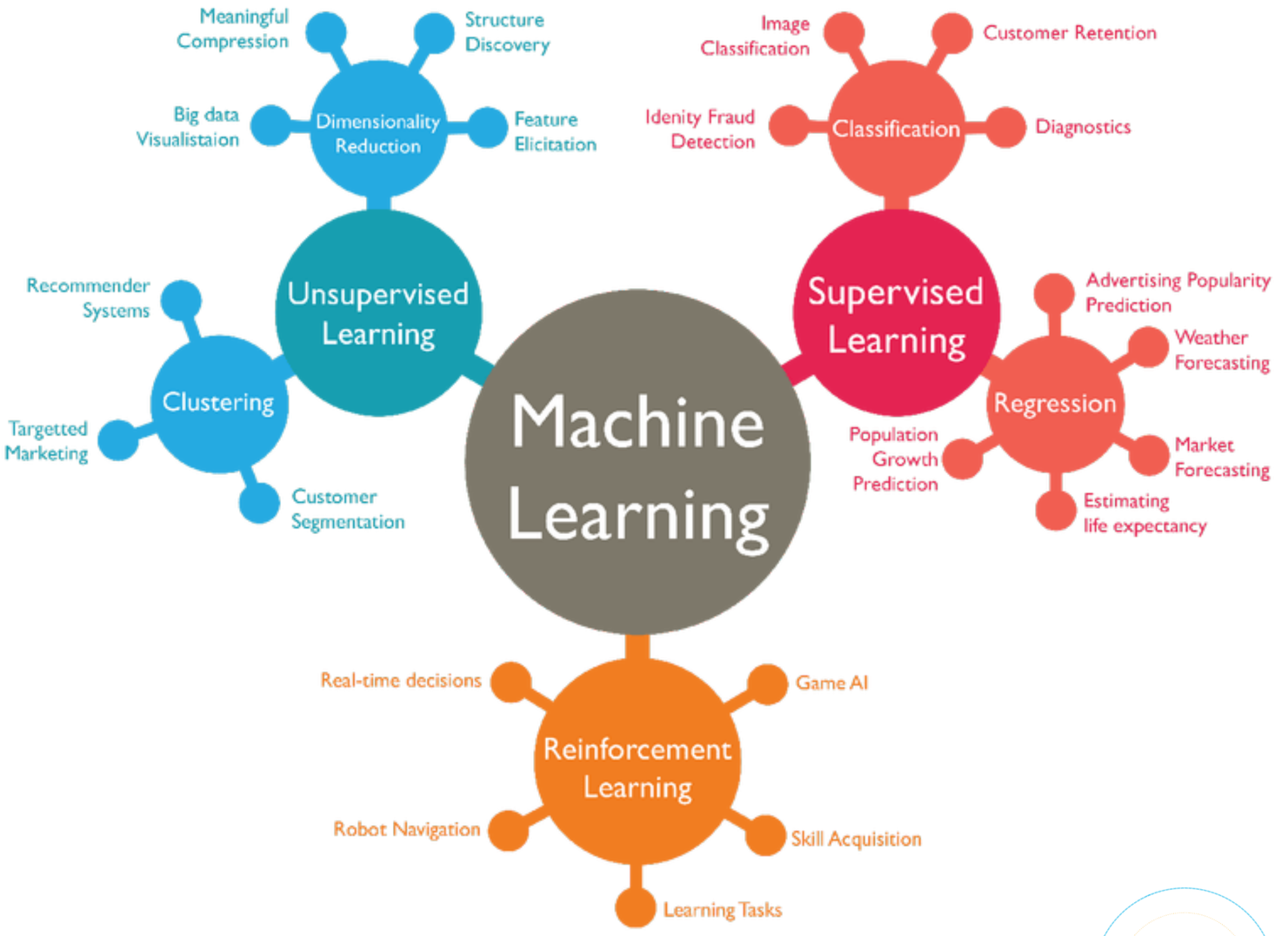
What is Machine Learning?



- Machine Learning
  - Study of algorithms that
  - improve their performance
  - at some task
  - with experience
- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

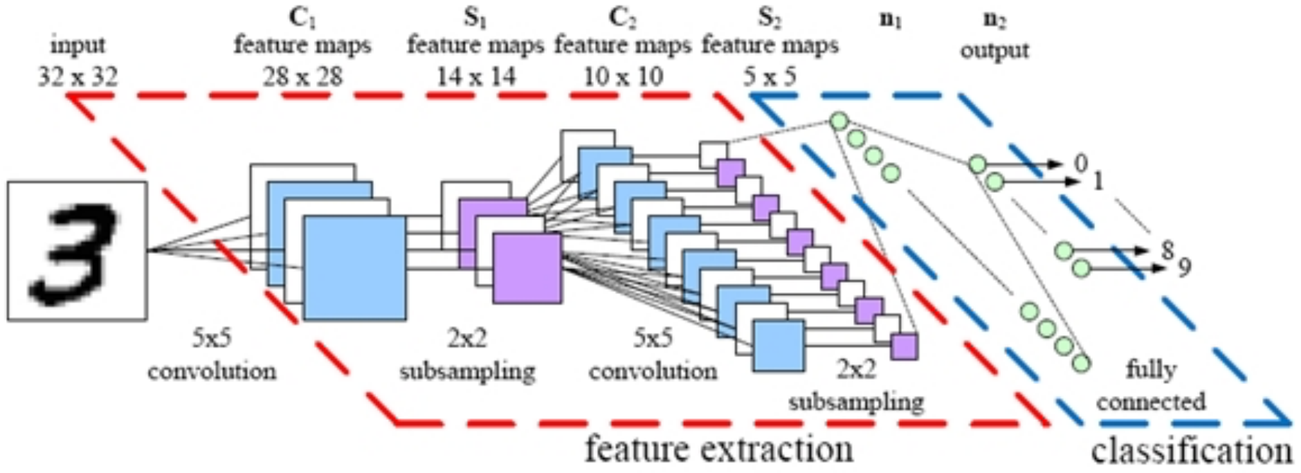
Applications

- Association Analysis
- Supervised Learning
  - Classification
  - Regression/Prediction
- Unsupervised Learning
- Reinforcement Learning



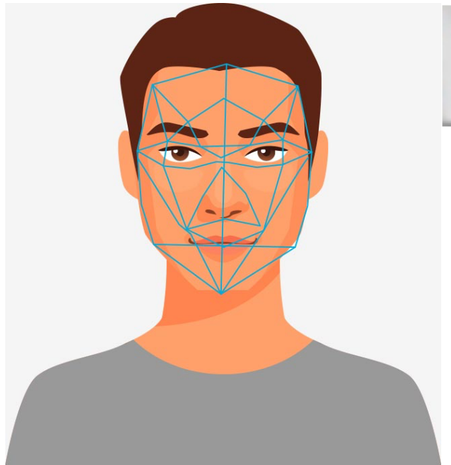
Classification

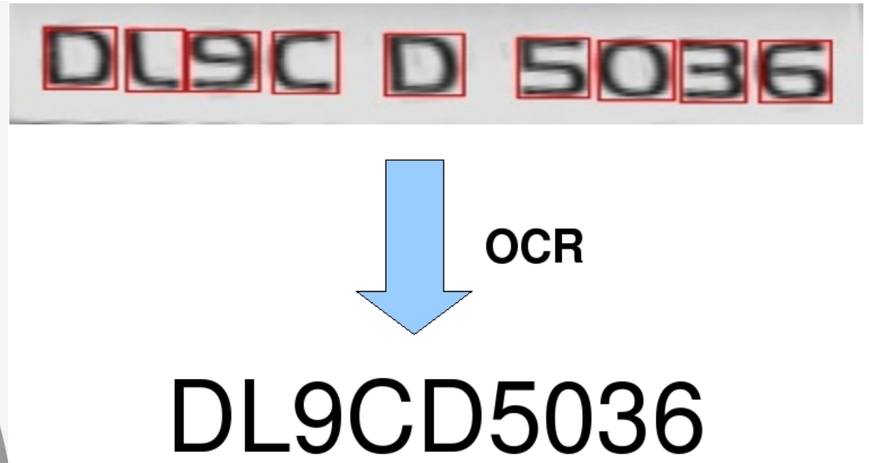
- 分类是监督学习，一般使用离散的类标（class label）
- 目的是对过往类标已知示例的观察与学习，实现对新样本类标的预测；
- 检测垃圾邮件的例子是典型的二类别分类（binary classification）任务



Classification: Applications

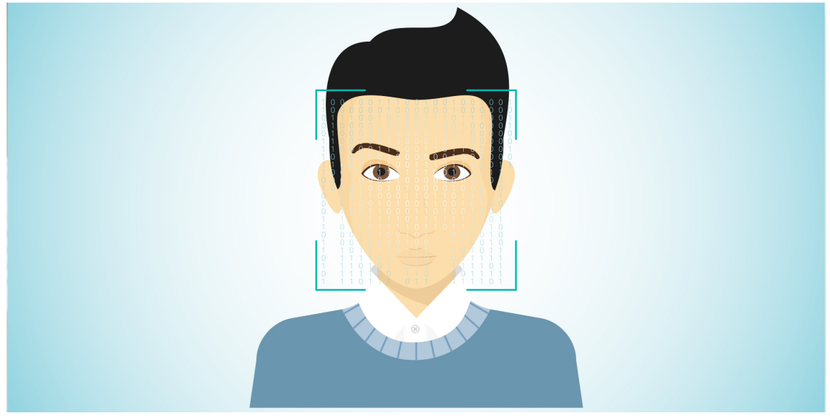
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- Web Advertising: Predict if a user clicks on an ad on the Internet.



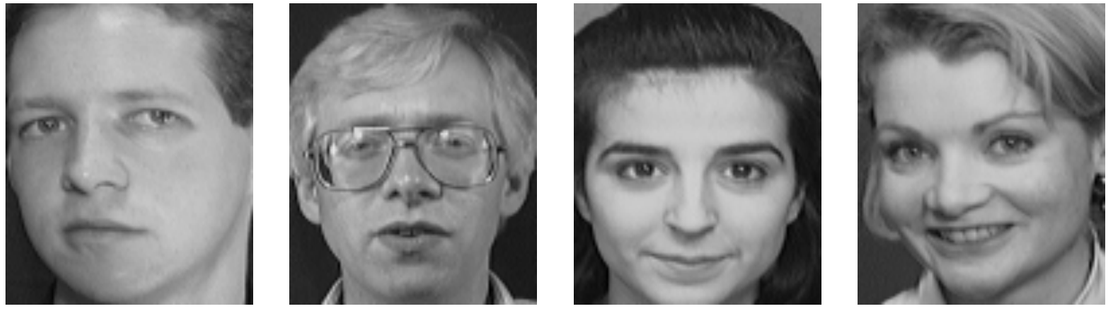


Face Recognition

Training examples of a person



Test images

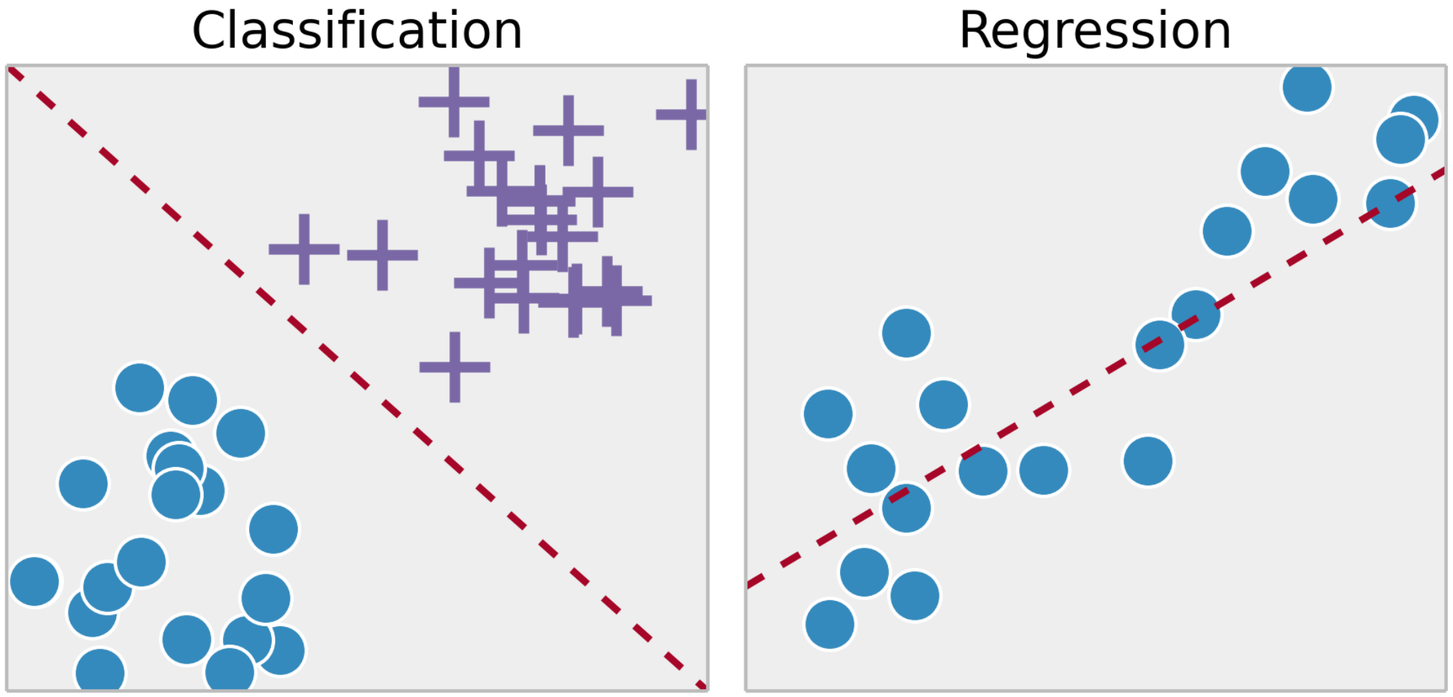


AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Prediction: Regression

Regression: re前缀表示的是恢复(recover),重复(repeat)的意思， sion尾缀把动词变成名词， gress的意思是行走

- 举个例子：
  预测明天的气温是多少度，这是一个回归任务；
  预测明天是阴、晴还是雨，就是一个分类任务。



Regression vs. Classification

分类和回归的区别主要在3个方面：

❑ 输出数据的类型

分类输出的数据类型是离散数据，也就是分类的标签。比如我们前面通过学生学习预测考试是否通过，这里的预测结果是考试通过，或者不通过，这2种离散数据。

回归输出是连续数据类型，比如我们通过学习时间预测学生的考试分数，这里的预测结果分数，是连续数据。

❑ 我们想要通过机器学习算法得到什么？

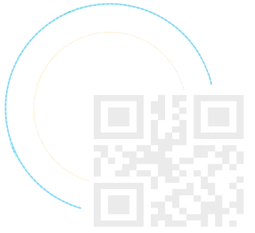
分类算法得到的是一个决策面，用于对数据集中的数据进行分类

回归算法得到的是一个最优拟合线，这个线条可以最好的接近数据集中的各个点

❑ 对模型的评估指标不一样

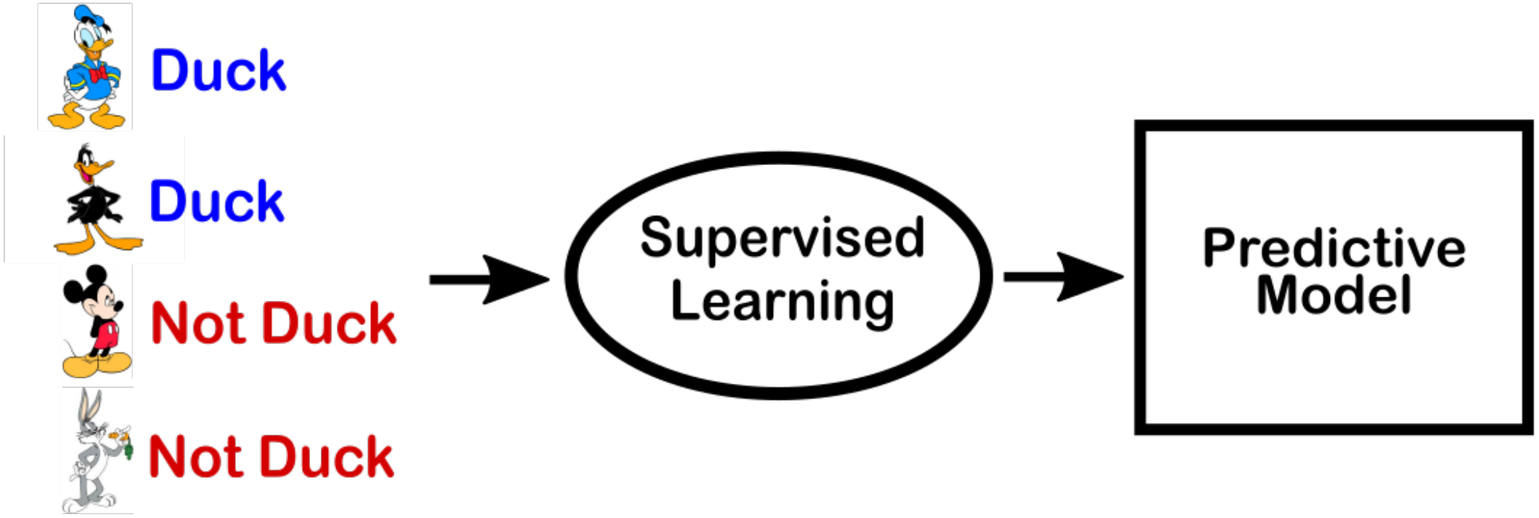
在监督分类中，我们通常会使用正确率作为指标，也就是预测结果中分类正确数据占数据的比例

在回归中，我们用决定系数R平方来评估模型的好坏。R平方表示有多少百分比的y波动被回归线描绘



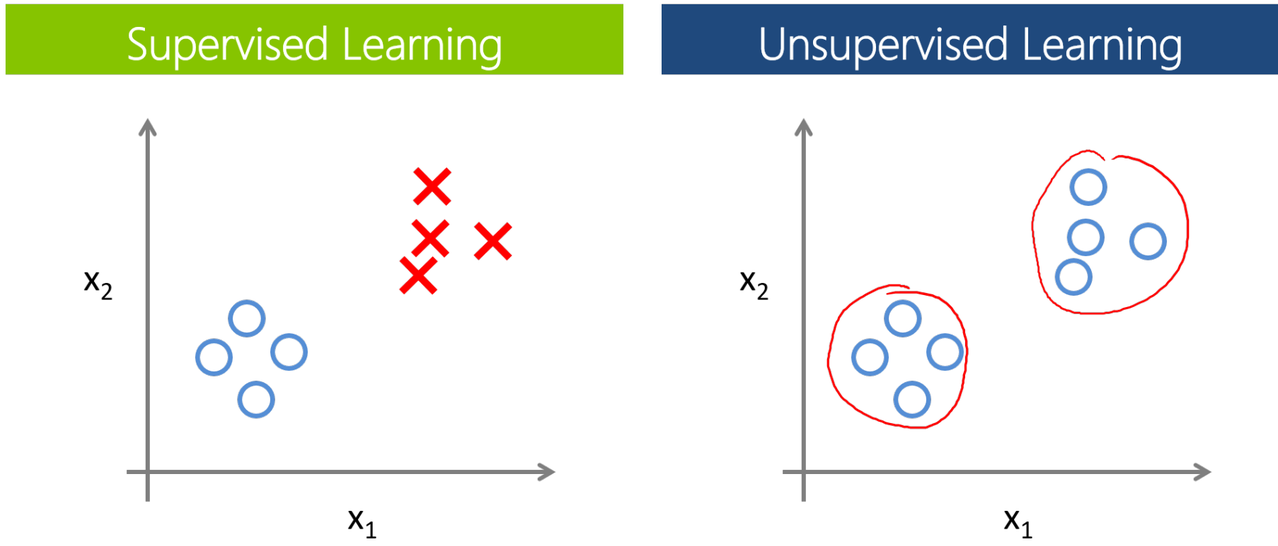
Supervised Learning: Uses

- Example: decision trees tools that create rules
  - Prediction of future cases: Use the rule to predict the output for future inputs
  - Knowledge extraction: The rule is easy to understand
  - Compression: The rule is simpler than the data it explains
  - Outlier detection: Exceptions that are not covered by the rule, e.g., fraud



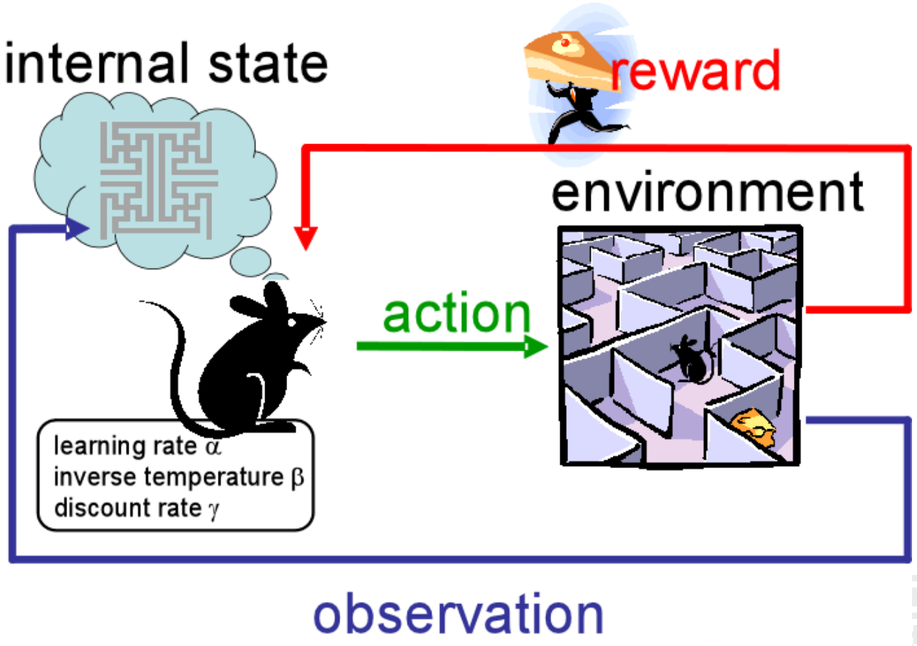
Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis
- Example applications
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs



Reinforcement Learning

- Topics:
  - Policies: what actions should an agent take in a particular situation
  - Utility estimation: how good is a state (used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...



Conclusion

- 下节课我们将学习自然语言基础

时间	课时安排
2018/2/6	第一课 NLP发展历史介绍和展望 1.NLP发展现状 2.传统NLP方法面临的挑战 3.Big Data和Deep Learning给NLP带来的变革和机遇 4.NLP的发展趋势，以及和各行各业的结合应用
2018/2/13	第二课 数学理论基础 1. 概率和信息论 2. 监督学习、半监督学习和非监督学习 3. 分类与回归模型
2018/2/20	第三课 自然语言基础 1. Word vector与Word embedding 2. 什么是分词、词性标注、依存句法分析等？如何利用开源工具包完成 3. 什么是统计自然语言处理？

END