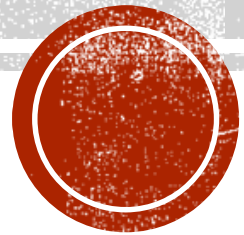
概率论简介

孙博士　七月在线



OUTLINE

- 1. 概率论基础
- 2. 随机变量和概率密度函数
  - 伯努利分布
  - 正态分布…
- 3. 信息论基础

符号简介

- X, x　　特征 (feature)
- Y, y　　标签 (label)
- D　　数据 (data)
- f　　模型 (model)
- θ　　参数 (parameter)
- L(X, Y, θ)　　目标函数 (objective function)
- $\min_\theta L(X, Y, \theta)$　　优化 (optimization)

概率

- **事件**发生可能的大小 （0~1）
  - 0 - 不可能发生 （直观上）
  - 1 – 一定会发生 （直观上）
  - 0.1 - 平均十次时间会发生一次

概率

- **事件**发生可能的大小 （0~1）
  - 0 - 不可能发生 （直观上）
  - 1 – 一定会发生 （直观上）
  - 0.1 - 平均十次时间会发生一次

- 机器学习的应用
  - 生成模型 vs 判别模型
  - 收到新的email, 判断邮件**是不是**垃圾邮件
  - 打开视频或者新闻网站，如何通过推荐提高用户点开链接的概率
  - 股票涨跌的概率
  - 图模型, 逻辑回归，决策树…
  - ….

事件

- 事件： 一个随机的过程的结果
  - 掷色子， 可能的结果 $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - 判断垃圾邮件， $\Omega = \{0, 1\}$
  - 判断推荐链接是否有效， $\Omega = \{0, 1\}$
  - 预测股票价格 $\Omega = [0, 100]$
  - $P(\Omega) = 1$
- 事件 ~ 集合

并集

- 如果A, B互斥， 那么$P(A\ or\ B) = P(A) + P(B)$ (or ~ ∪)
  - 是 {1,2} 或者 {3,4} 的概率是 1/3 + 1/3 = 2/3
  - 是 {1,2} 或者 {2,3} 的概率不是 1/3 + 1/3

并集

- 如果A, B互斥， 那么$P(A\ or\ B) = P(A) + P(B)$ (or ~ ∪)
  - 是 {1,2} 或者 {3,4} 的概率是 1/3 + 1/3 = 2/3
  - 是 {1,2} 或者 {2,3} 的概率不是 1/3 + 1/3

- $P(A) + P(A^c) = 1$
  - 色子结果 是3 和 不是3 的概率和为1

并集

- 如果A, B互斥， 那么$P(A\ or\ B) = P(A) + P(B)$ (or ~ ∪)
  - 是 {1,2} 或者 {3,4} 的概率是 1/3 + 1/3 = 2/3
  - 是 {1,2} 或者 {2,3} 的概率不是 1/3 + 1/3

- $P(A) + P(A^c) = 1$
  - 色子结果 是3 和 不是3 的概率和为1

- 假设一共有n种互斥的可能, 那么 $P(\Omega) = \sum_{i=1}^{n} P(A_i) = 1$

独立事件

- $P(A \cap B) = P(A) * P(B)$那么A和B是独立的
  - 直观： A的发生与否对B没有影响

独立事件

- $P(A \cap B) = P(A) * P(B)$那么A和B是独立的
  - 直观： A的发生与否对B没有影响
  - EX1： 掷三次硬币，正反反的概率是1/2 * 1/2 * 1/2
  - EX2： 收到三封邮件，都是垃圾邮件的概率是 0.01*0.01*0.01
  - EX3： 收到三封来自同一个邮箱的邮件， 都是垃圾邮件的概率?

独立事件

- $P(A \cap B) = P(A) * P(B)$那么A和B是独立的
  - 直观： A的发生与否对B没有影响
  - EX1： 掷三次硬币，正反反的概率是1/2 * 1/2 * 1/2
  - EX2： 收到三封邮件，都是垃圾邮件的概率是 0.01*0.01*0.01
  - EX3： 收到三封来自同一个邮箱的邮件， 都是垃圾邮件的概率?
  - 大部分情况下，机器学习模型假设**数据**是**独立**的

条件概率

- P(Y|X)
  - 直观上， 在X发生的情况下， 发生Y的概率

条件概率

- P(Y|X)
  - 直观上， 在X发生的情况下， 发生Y的概率
  - EX1: 已知邮件X=jd或者taobao, 那么邮件是推销邮件的概率
  - EX2: 已知文章标题里有“震惊！快转！愤怒！”， 那么用户点击的概率

条件概率

- P(Y|X)
  - 直观上， 在X发生的情况下， 发生Y的概率
  - EX1: 已知邮件X=jd或者taobao, 那么邮件是推销邮件的概率
  - EX2: 已知文章标题里有“震惊！快转！愤怒！”， 那么用户点击的概率
  - 机器学习模型！

联合概率

- $P(XY) = P(X) * P(Y|X)$
  - 直观上，X和Y同时发生 = X先发生， X发生的情况下Y发生
  - 如果X,Y独立， 那么P(Y|X) = P(Y)
  - P(Y|X) = P(XY)/P(X)

联合概率

- $P(XY) = P(X) * P(Y|X)$
  - 直观上，X和Y同时发生 = X先发生， X发生的情况下Y发生
  - 如果X,Y独立， 那么P(Y|X) = P(Y)
  - P(Y|X) = P(XY)/P(X)

- $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \dots P(X_n|X_{n-1}, X_{n-2}, \dots, X_1)$
  - 马尔科夫(Markov) $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2\ ) \dots P(X_n|X_{n-1})$

联合概率

- $P(XY) = P(X) * P(Y|X)$
  - 直观上，X和Y同时发生 = X先发生， X发生的情况下Y发生
  - 如果X,Y独立， 那么P(Y|X) = P(Y)
  - P(Y|X) = P(XY)/P(X)

- $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \dots P(X_n|X_{n-1}, X_{n-2}, \dots, X_1)$
  - 马尔科夫(Markov) $P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2) \dots P(X_n|X_{n-1})$

- $P(X) = \sum_i P(X|Y_i) * \mathrm{P}(\mathrm{Y_i})$， Y_i不相交

贝叶斯公式

- $P(XY) = P(X|Y)*P(Y) = P(Y|X)*P(X)$

贝叶斯公式

- P(XY) = P(X|Y)*P(Y) = P(Y|X)*P(X)

- $P(Y|X) = P(X|Y) * P(Y) / P(X)$
  - P(Y|X) 后验概率 （posterior)
  - P(Y) 先验概率 (prior)

贝叶斯公式

- P(XY) = P(X|Y)*P(Y) = P(Y|X)*P(X)

- $P(Y|X) = P(X|Y) * P(Y) / P(X)$
  - P(Y|X) 后验概率 （posterior)
  - P(Y) 先验概率 (prior)

- EX1: 含有sex的邮件是垃圾邮件的概率 P(Y=spam|X=sex)
  - P(Y=spam)=0.9 先验
  - 假设垃圾邮件中出现sex的概率是1%， 正常邮件中出现的概率是 0.01%
  - P(Y=spam|X=sex) = P(X=sex|Y=spam)*P(Y=spam)/P(X=sex)
  - 计算上面公式! （hw1,提示， P(X=sex)分情况讨论)

生成模型和判别模型

- 目标 $P(Y|X)$

生成模型和判别模型

- 目标 P(Y|X)

- 生成模型 $P(Y|X) = \mathrm{P(X|Y)} * \mathrm{P(Y)} / \mathrm{P(X)}$
  - 朴素贝叶斯 (Naïve Bayes)
  - 隐马尔科夫 (Hidden Markov Model)

生成模型和判别模型

- 目标 P(Y|X)

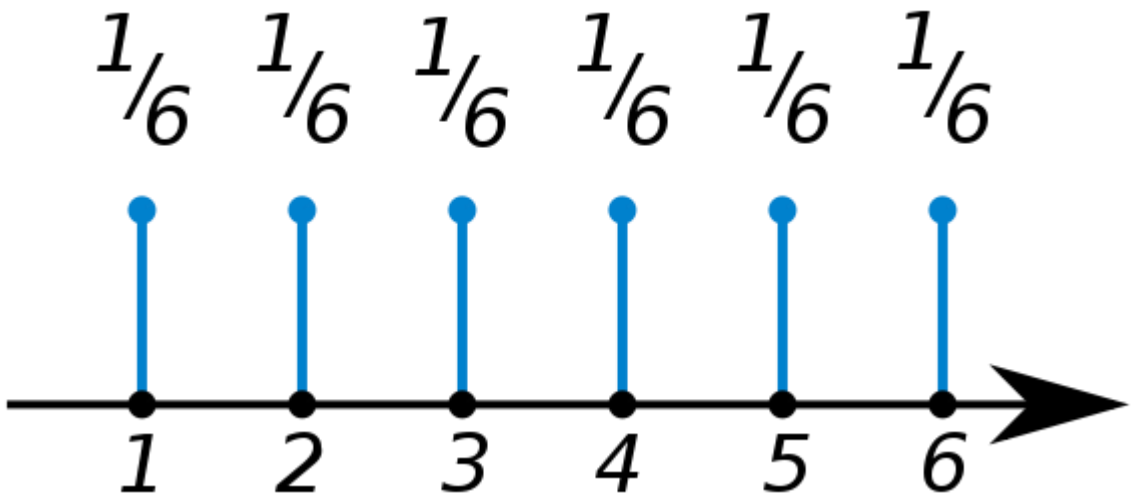
- 生成模型 $P(Y|X) = P(X|Y) * P(Y) / P(X)$
  - 朴素贝叶斯 (Naïve Bayes)
  - 隐马尔科夫 (Hidden Markov Model)

- 判别模型 P(Y|X)
  - 逻辑回归 (Logistic Regression)
  - 支持向量机 (Support Vector Machine)
  - 条件随机场 (Conditional Random Field)
  - …

随机变量

- X: 集合到实数的映射
  - 色子： 1到6
  - 硬币： 正面~0， 反面~1
  - 股票价格: 价格
  - 垃圾邮件： 是~1， 否~0
  - 分为离散的连续的

离散随机变量



Probability Mass Function
随机质量函数

离散随机变量

- 伯努利分布(Bernoulli distribution): P(Y=1) = **p** = 1-P(Y=0)=1-q
- 分类分布(categorical distribution)： 多个离散值, 参数是(P1, P2, …)

离散随机变量

- 伯努利分布(Bernoulli distribution): P(Y=1) = **p** = 1-P(Y=0)=1-q
- 分类分布(categorical distribution)：多个离散值, 参数是(P1, P2, …)

- 二项分布 (binomial distribution): n次伯努利， k次成功的概率
- 多项分布 (multinomial distribution)： n次分类分布,(k1, k2, ..kn)的概率

期望

- 假设x1, x2, …, xn对应的概率为p1, p2, …, pn, 那么X的期望(Expectation)为

$$E[X] = x_1 p_1 + x_2 p_2 + \cdots x_n p_n$$

期望

- 假设x1, x2, …, xn对应的概率为p1, p2, …, pn, 那么X的期望(Expectation)为

$$E[X] = x_1p_1 + x_2p_2 + \cdots x_np_n$$

- 直观上，是随机变量X的“平均数”
- 性质：
  - E[X+Y]=E[X]+E[Y], E[ax] = aE[X]
  - 如果X, Y独立，那么 E[XY]=E[X]*E[Y]

期望

- 假设x1, x2, …, xn对应的概率为p1, p2, …, pn, 那么X的期望(Expectation)为

$$E[X] = x_1p_1 + x_2p_2 + \cdots x_np_n$$

- 直观上，是随机变量X的“平均数”
- 性质：
  - E[X+Y]=E[X]+E[Y], E[ax] = aE[X]
  - 如果X, Y独立，那么 E[XY]=E[X]*E[Y]

- 伯努利分布的期望： p
- 二项分布的期望： np （hw2, 提示，利用期望的性质）

方差

- 假设μ为期望， x1, x2, …, xn对应的概率为p1, p2, …, pn, 那么X的方差(Variance)为

$$Var[X] = (x_1 - \mu)^2 p_1 + \cdots + (x_n - \mu)^2 p_n$$

- Var[X] = $E[(X - \mu)^2]$

方差

- 假设μ为期望， x1, x2, …, xn对应的概率为p1, p2, …, pn, 那么X的方差(Variance)为

$$Var[X] = (x_1 - \mu)^2 p_1 + \cdots + (x_n - \mu)^2 p_n$$

- Var[X] = $E[(X - \mu)^2]$
- 离平均数的平均偏离值
- 性质:
  - $Var[X] \geq 0$
  - $Var[X] = E[X^2] - E[X]^2$
  - $Var[X + a] = Var[X],\ Var[aX] = a^2 Var[X]$
  - 如果X和Y独立， Var[X+Y]=Var[X]+Var[Y]

方差

- 假设μ为期望， x1, x2, …, xn对应的概率为p1, p2, …, pn, 那么X的方差(Variance)为

$$Var[X] = (x_1 - \mu)^2 p_1 + \cdots + (x_n - \mu)^2 p_n$$

- Var[X] = $E[(X - \mu)^2]$
- 离平均数的平均偏离值
- 性质：
  - $Var[X] \geq 0$
  - $Var[X] = E[X^2] - E[X]^2$
  - $Var[X + a] = Var[X],\ Var[aX] = a^2 Var[X]$
  - 如果X和Y独立， Var[X+Y]=Var[X]+Var[Y]

- 伯努利分布的方差： p*(1-p)
- 二项分布的方差： np*(1-p) (hw3, 提示，利用方差的性质)

机器学习实例-ROC曲线介绍

- 假设某个分类模型f, 有n个邮件$x_1, x_2, \ldots, x_{100}$
- 模型预测垃圾邮件的概率为 $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{100}$, 真实的label为$y_1, y_2, \ldots, y_{100}$
- 思考: 如何评估模型的表现?

机器学习实例-ROC曲线介绍

- 假设某个分类模型f, 有n个邮件$x_1, x_2, \ldots, x_{100}$
- 模型预测垃圾邮件的概率为 $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{100}$, 真实的label为$y_1, y_2, \ldots, y_{100}$
- 思考: 如何评估模型的表现?
  - 准确率， 有什么缺陷?

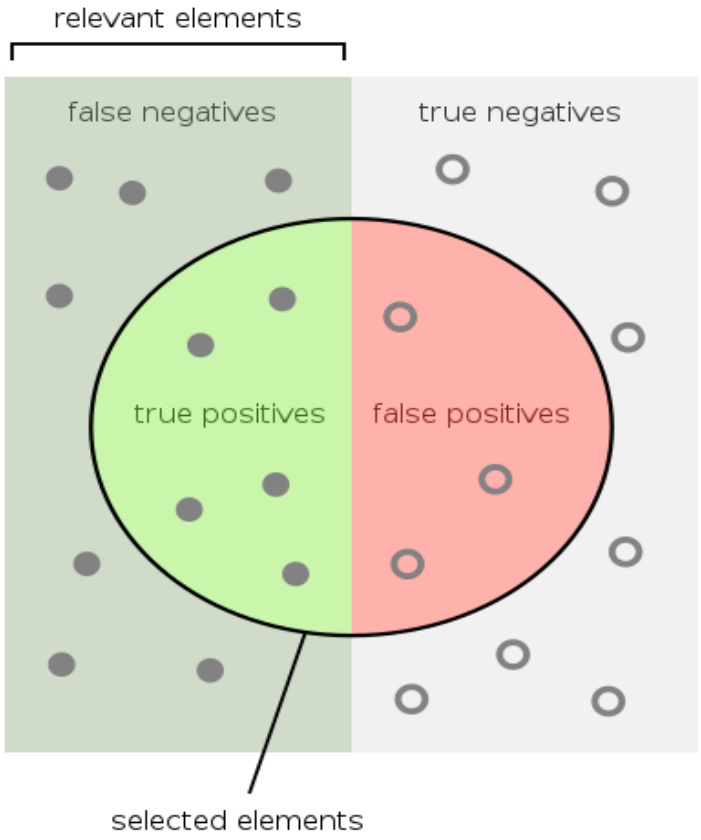
机器学习实例-ROC曲线介绍

- 假设某个分类模型f, 有n个邮件$x_1, x_2, \ldots, x_{100}$
- 模型预测垃圾邮件的概率为 $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{100}$, 真实的label为$y_1, y_2, \ldots, y_{100}$
- 思考: 如何评估模型的表现?
  - 准确率， 有什么缺陷?
  - 假设数据label不平衡， 0-1比例为 99% vs 1%

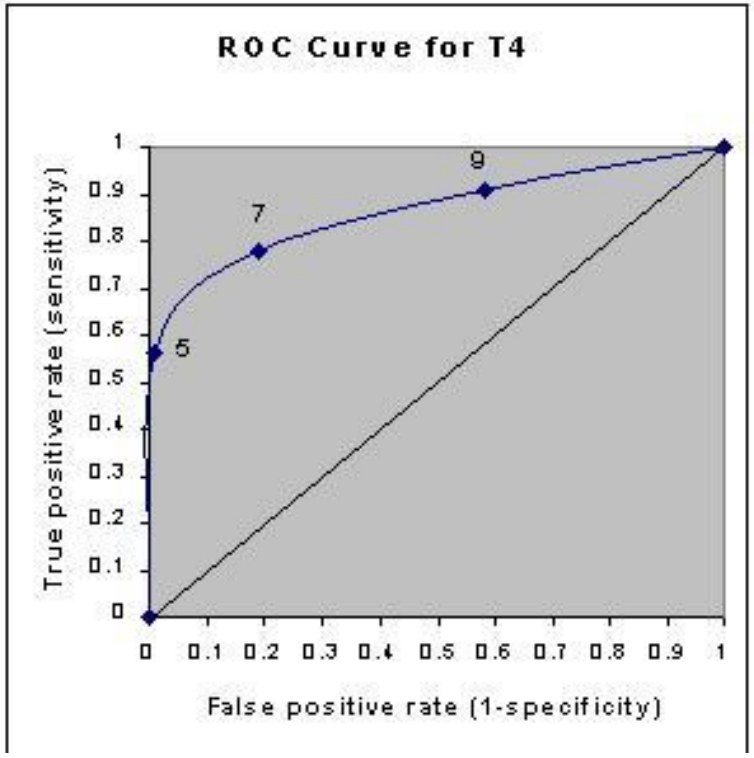
机器学习实例-ROC曲线介绍

- 假设某个分类模型f, 有n个邮件$x_1, x_2, \ldots, x_{100}$
- 模型预测垃圾邮件的概率为 $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{100}$, 真实的label为$y_1, y_2, \ldots, y_{100}$
- 思考: 如何评估模型的表现?
  - 准确率， 有什么缺陷?
  - 假设数据label不平衡， 0-1比例为 99% vs 1%
  - 如何选择最有可能的垃圾邮件?
  - 推荐系统的相似性

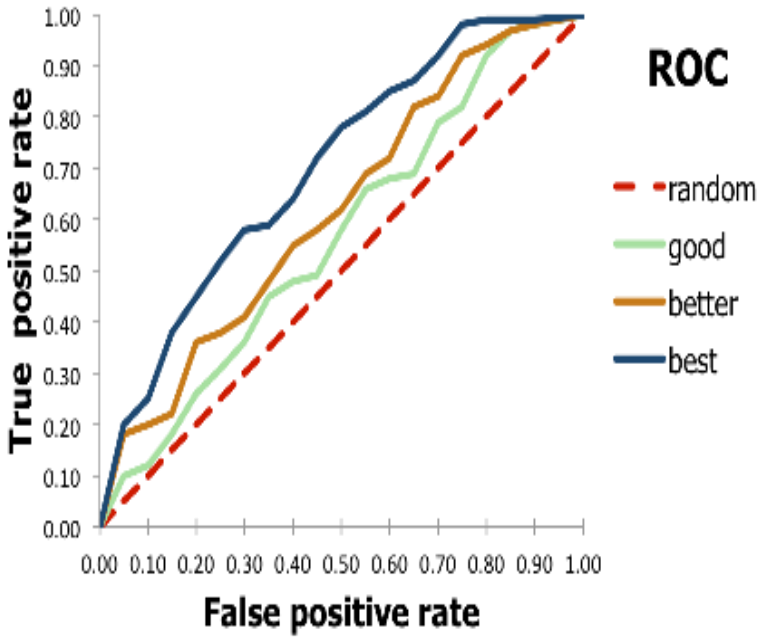
机器学习实例-ROC曲线定义

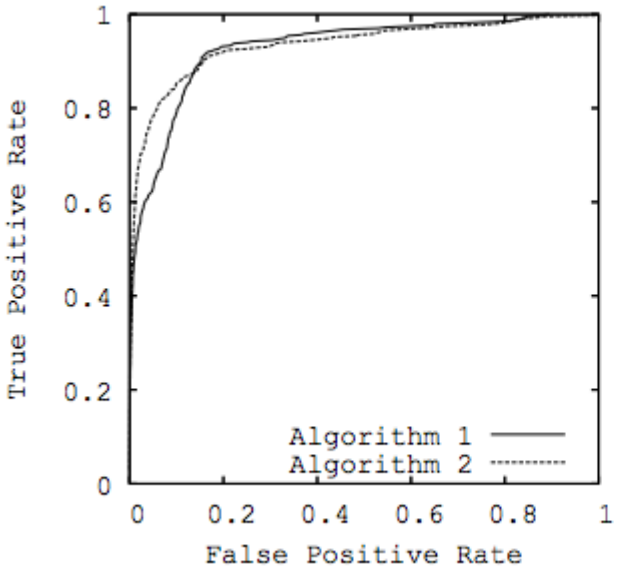


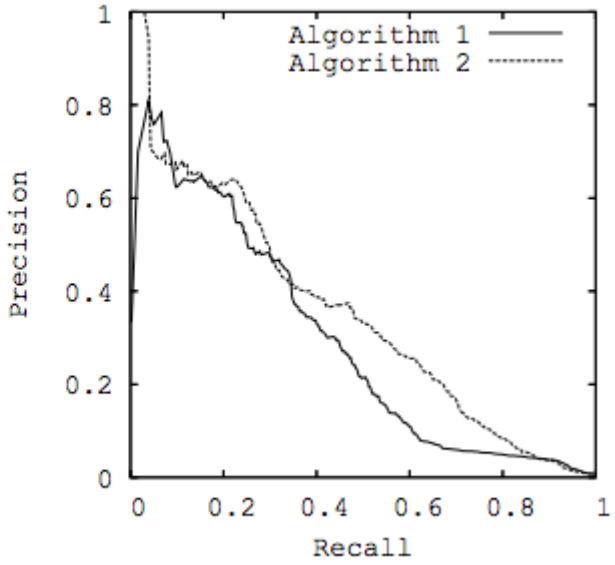
- precision = TP / (TP+FP)
- **recall = TPR = TP/(TP+FN)**
- **FPR = FP/(FP+TN)**



机器学习实例-ROC曲线评价







Ref: The Relationship Between Precision-Recall and ROC Curves

连续随机变量

- 条件： $f(X) \geq 0,\ X \in \Omega, (f(X) \leq 1\ ?), \int f(x)dx = 1$
- 概率： $P(X \in S) = \int_S f(x)dx$
- 期望： $E[X] = \int Xf(X)dx$
- 方差： $Var[X] = \int (X-\mu)^2 f(X)dx$

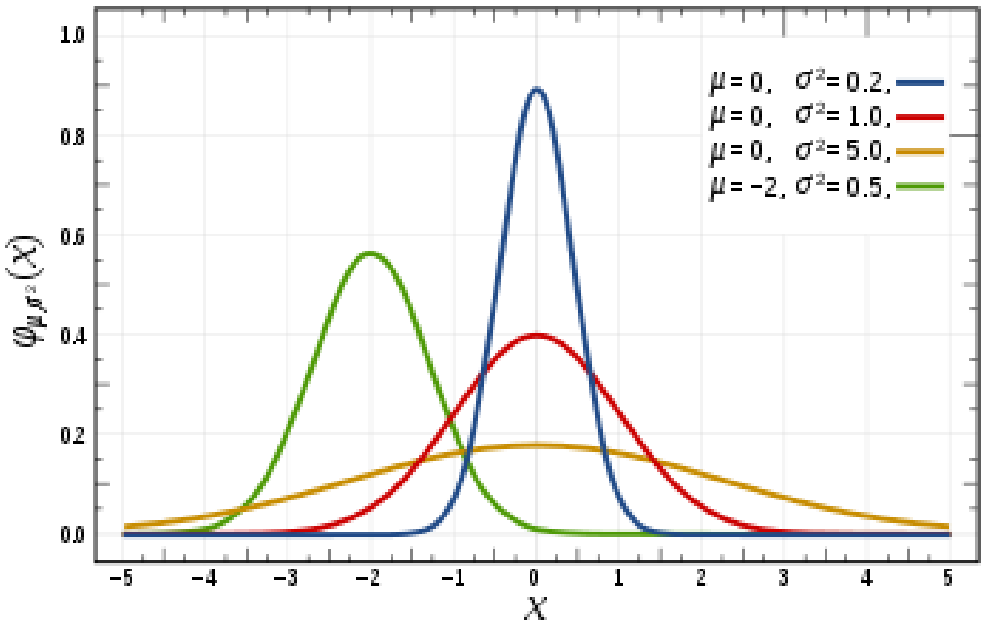
连续随机变量

- 条件： $f(X) \geq 0,\ X \in \Omega, (f(X) \leq 1\ ?), \int f(x)dx = 1$
- 概率： $P(X \in S) = \int_S f(x)dx$
- 期望： $E[X] = \int Xf(X)dx$
- 方差： $Var[X] = \int (X-\mu)^2 f(X)dx$

- 常见问题：
  - P(X)=0 一定是不可能发生的事件吗？ (P(X)=1)
  - 期望一定存在吗？ (large tails, 柯西分布)

正态分布

- $X \sim N(\mu, \sigma^2),\ f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$



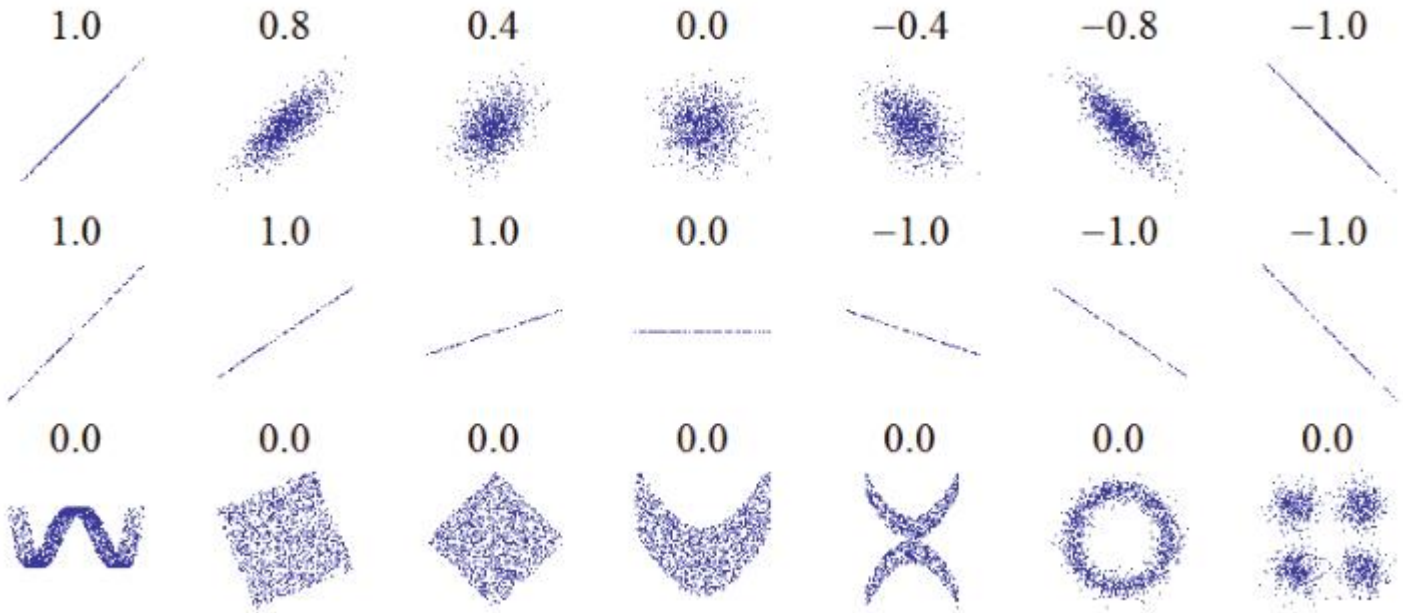
$E[X] = \mu,$ $\mathrm{Var}[X] = \sigma^2$

协方差和相关系数

- $COV(X,Y) = E[\,(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- $COR(X,Y) = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (hw4, 证明|cor|<=1, 提示 Cauchy–Schwarz不等式)

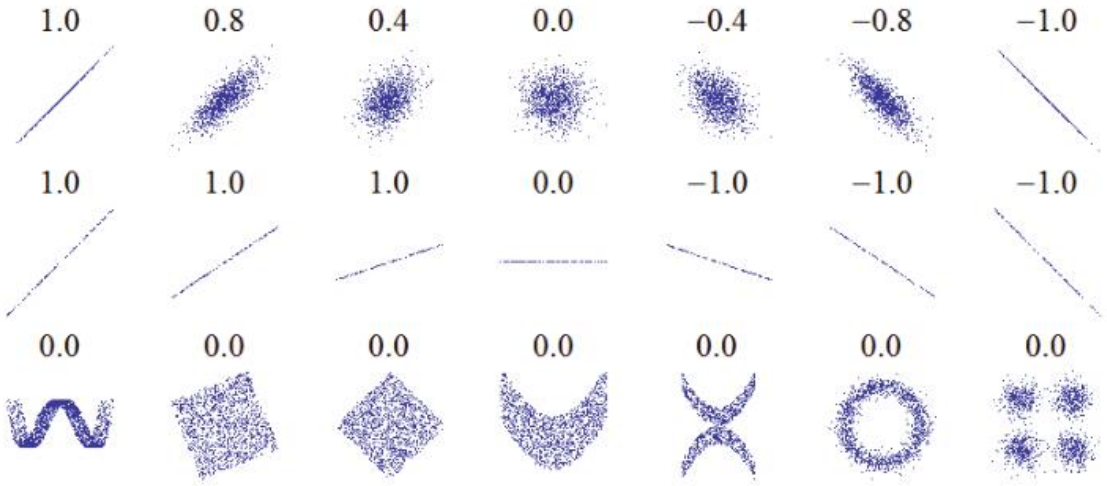
协方差和相关系数

- $COV(X,Y) = E[\,(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- $COR(X,Y) = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (hw4, 证明|cor|<=1, 提示 Cauchy–Schwarz不等式)



协方差和相关系数

- $COV(X,Y) = E[\,(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- $COR(X,Y) = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (hw4, 证明|cor|<=1, 提示 Cauchy–Schwarz不等式)
- Var[X+Y] = Var[X]+Var[Y] + COV(X,Y)



协方差矩阵

$x = (X_1, X_2, \ldots, X_d)^T$

$$\begin{aligned}\operatorname{cov}[\mathbf{x}] &\triangleq \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \\ &= \begin{pmatrix} \operatorname{var}[X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_d] \\ \operatorname{cov}[X_2, X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_d, X_1] & \operatorname{cov}[X_d, X_2] & \cdots & \operatorname{var}[X_d] \end{pmatrix}\end{aligned}$$

多元高斯

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)]$$

中心极限定理

- 大数定理
  - $X_1, X_2, \ldots, X_n,\ i.i.d, E[X_i] = \mu$,那么 $\bar{X}_n = \frac{(X_1 + \cdots + X_n)}{n} \to \mu$ 当$n \to \infty$
- 中心极限定理
  - $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{N}} \sim N(0,1)$
- MLE (simulation)

蒙特卡洛近似

- $E[f(X)] = \int f(x)P(x)dx \approx \frac{1}{N}\sum_i f(X_i)$

- 通过改变f(X), 可以得到
  - $\bar{x} = \frac{1}{N}\sum_i x_i \ \rightarrow E[X]$
  - $\frac{1}{N}\sum_i (x_i - \bar{x})^2 \ \rightarrow Var[X]$
  - $\frac{1}{N}\#\{x_i \leq c\} \ \rightarrow \ P(X \leq c)$

机器学习实例-NAÏVE BAYES

- $P(Y|X_1, X_2, \dots, X_p) = \frac{P(X_1, X_2, .., X_p|Y)P(Y)}{P(X_1, X_2, \dots, X_p)} = \frac{P(X_1|Y)P(X_2|Y)\dots P(X_p|Y)P(Y)}{P(X_1, X_2, \dots, X_p)}$
- $\arg\max_k P(Y = k|X_1, X_2, \dots, X_p)$ **(MAP)**

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Ref: Wikipedia (Naive_Bayes_classifier)

机器学习实例-NAÏVE BAYES

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033*10-02	176.25	1.2292*10+02	11.25	9.1667*10-01
female	5.4175	9.7225*10-02	132.5	5.5833*10+02	7.5	1.6667

机器学习实例-NAÏVE BAYES

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033*10-02	176.25	1.2292*10+02	11.25	9.1667*10-01
female	5.4175	9.7225*10-02	132.5	5.5833*10+02	7.5	1.6667

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

机器学习实例-NAÏVE BAYES

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033*10-02	176.25	1.2292*10+02	11.25	9.1667*10-01
female	5.4175	9.7225*10-02	132.5	5.5833*10+02	7.5	1.6667

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

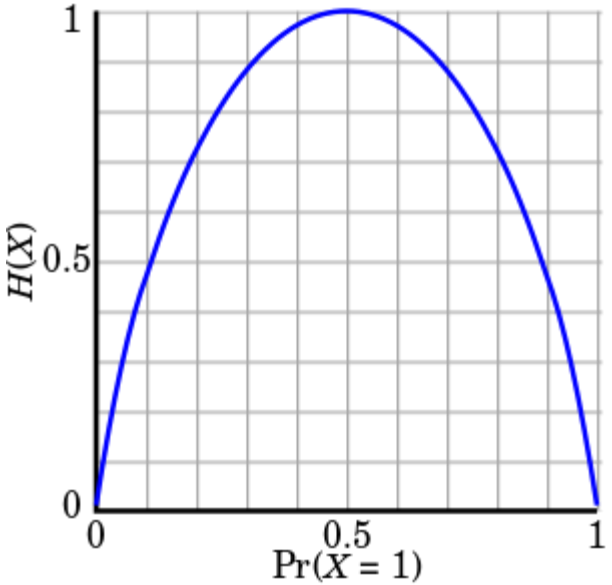
$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789$$

Post(female) ~ 10^{-4}
Post(male) ~ 10^{-9}

熵(ENTROPY)

- $H(X) = -\sum_i P(X_i) \log P(X_i)$ $\quad (E_P[-\log P(X)])$
- 代表不确定性!
- EX： 对于伯努利分布 H(X)= -P*logP-(1-P)log(1-P)



机器学习应用-决策树

- $H(X) = -\sum_i P(X_i) \log P(X_i)$
- 代表不确定性!
- EX： 对于伯努利分布 H(X)= -P*logP-(1-P)log(1-P)
- [9男, 5女] → Entropy = - (5/14)*log(5/14) - (9/14)*log(9/14) = 0.9403
- 名字是否有 “雨”
  - 有 [3男， 4女], Entropy1 = - (3/7)*log(3/7) - (4/7)*log(4/7) = 0.9852
  - 没有[6男, 1女], Entropy2 = - (6/7)*log(6/7) - (1/7)*log(1/7) = 0.5917
  - Entropy_new = 7/14*Entropy1 + 7/14*Entropy2 = 0.7885
  - Information_Gain = Entropy - Entropy_new = 0.1518

KL DIVERGENCE

- 给定两个概率分布p, q， 定义KL Divergence为

$$KL(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

- 或者$KL(p||q) = \sum_i p_i \log p_i - \sum_i p_i \log q_i = -H(p) + H(p,q)$

- $KL(p||q) \geq 0$, 当且仅当p==q时， KL(p||q)=0 (hw5, 提示， Jensen inequality)

- KL Divergence不对称!

- 常用于解释EM算法

互信息(MUTUAL INFORMATION)

- $I(X,Y) = KL(P(X,Y)||P(X)P(Y)) = \sum_x \sum_y P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}$
- $I(X,Y) \geq 0$, 当且仅当$P(X,Y) = P(X)P(Y)$时， $\mathrm{I(X,Y)=0}$
- $\mathrm{I(X,Y) = H(X) - H(X|Y)}$

谢谢！