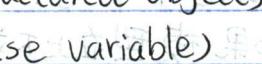
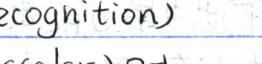
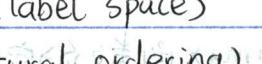


Machine

Learning - P1

监督学习 (supervised learning), 也称预测性学习 (predictive learning)
目的是学习一个从输入 X 到输出 y 的映射 (mapping) 
 $D = \{(X_i, y_i)\}_{i=1}^N$ 为一个 labeled set of input-output pairs
称为训练集 (training set) 
输入 X_i 的简单情形是一个多维向量 (n -dimensional vector) of numbers
称为特征 (feature) / 属性 (attribute) / 协变量 (covariates)
一般为更复杂的对象 (complex structured object)
对于输出 y_i , 也称为因变量 (response variable) 
当 $y_i \in \{1, 2, \dots, C\}$ 时,
称为分类变量 (categorical variable) / nominal variable
大部分模型都会假设输出是分类变量
通常为分类问题 (classification) 
或模式识别 (pattern recognition) 
当 y_i 为实数标量 (real-valued scalar) 时
通常为回归问题 (regression) 
当 $y_i \in Y$, 其中 Y 是命名空间 (label space)
且 Y 服从某种自然排序 (natural ordering) 
通常为有序回归 (ordinal regression) 

非监督学习 (unsupervised learning), 也称描述性学习 (descriptive learning)

与监督学习相比, 学习的仅有输入集合 $D = \{X_i\}_{i=1}^N$

目的是找到“感兴趣的模式 (interesting pattern)”

通常为知识发现问题 (knowledge discovery)

非良定义的问题 (much less well-defined problem)

没有明显的可使用的误差度量

强化学习 (reinforcement learning, RL)

对于给定的偶然的奖励/惩罚信号

occasional reward / punishment signal

学习如何行动

learn how to act / behave

Contextual reinforcement learning: 在特定上下文中学习如何行动

Machine

Learning - P₂

对给定的输入向量 (input vector) \vec{x} 和训练集 (training set) D

如果输出 y 是一个分类变量, $y \in \{1, 2, \dots, c\}$

则有 distribution of probability over possible label

$$p(y|\vec{x}, D)$$

即给定条件 \vec{x} 和 D 的 y 的条件概率

当选择不同的模型 (model) 时

也会记作 $p(y|\vec{x}, D, M)$, 其中 M 表示选择的模型

如果模型在上下文中是明确的, 也可以省略 M 而记为 $p(y|\vec{x}, D)$

model is clear from context

最优猜测 (best guess, true label), 对于给定的概率输出 (probabilistic output),

$$\hat{y} = \hat{f}(\vec{x}) = \arg \max_{c=1}^C p(y=c|\vec{x}, D)$$

为输出中最可能出现的分类标签 (most probable class (label))

称为分布 $p(y|\vec{x}, D)$ 的模 (mode) (MAP estimation)

而这个过程为最大后验概率估计 (maximum a posteriori estimation,

其中 $\arg \max$ 函数的定义为, 对于给定的函数 $f: X \rightarrow Y$

$$\arg \max_{x \in S \subseteq X} f(x) = \{x \mid x \in S \wedge \forall y \in S \quad f(y) \leq f(x)\}$$

即描述为使 $f(x)$ 取到最大值的 ~~所有点~~ ~~所有可能的点~~ ~~所有可能的点~~ x 的点集

其中 S 为给定的 x 的取值范围

通常指定 $x \in S$ 表示并非在 $f(x)$ 的整个定义域 X 上搜索, 即 $S \subset X$

当 S 平凡地为 X 或根据上下文可以明确角 S 时

$$\arg \max_x f(x) = \{x \mid \forall y \quad f(y) \leq f(x)\}$$

相比于最大值函数 $\max f(x) = \{f(x) \mid \forall y \quad f(y) \leq f(x)\}$

向量范数 (norm), 对于 n 维欧几里得空间 (Euclidean space) R^n 上的向量 $\vec{x} = (x_1, x_2, \dots, x_n)$

$$P \text{ 范数 } (P-norm) \text{ 定义为 } \|\vec{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \text{ 其中 } p \in N^+$$

注意当 $p=2$ 时, 称为欧几里得范数 (Euclidean norm)

当 $\vec{x} = (x_1, x_2, \dots, x_n)$ 为列向量 (column vector) 时

$$\text{有 } \|\vec{x}\|_2 = \sqrt{\vec{x}^T \vec{x}}$$

另外有等价于点乘, 即 $\|\vec{x}\|_2 = \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{\sum_{i=1}^n x_i^2}$

当 $p=1$ 时, 称为出租车范数 (Taxicab norm) 或曼哈顿范数 (Manhattan norm)

称 $\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$ 为曼哈顿距离 (Manhattan distance)

Machine

Learning - P3

参数模型 (parametric model), probabilistic model of form $p(y|\vec{x})$

具有固定数量的参数 (fixed number of parameters)

优点是速度更快 (being faster to use)

缺点是对于数据本身的分布有着强假设

making stronger assumption about nature of data distribution

- 类似于通过结构化表达式和参数表的模型

参数机器学习算法：简化目标函数为已知形式的算法

通过固定大小的参数集拟合数据学习模型

参数集与训练样本数独立，不论输入多少数据，对其需要的参数数量没有影响

包含两部分：选择目标函数形式

从训练数据中学习目标函数的参数

目标函数的形式通常假设为对于输入变量的线性关系

于是也称为线性机器学习算法

优势：简洁：容易理解理论和解释结果

快速：参数模型学习和训练的速度都很快

数据更少：通常不需要大量数据，在对数据拟合不好时表现也不错

局限性：约束：选定的函数形式限制了模型

有限的复杂度：通常只能应对简单的问题

拟合度小：实际中通常无法与潜在的目标函数拟合

非参数模型 (non-parametric model), probabilistic model of form $p(\vec{x})$

参数数量随着训练集增加而增加 (number of parameters grow with amount of training data)

优点是更灵活 (more flexible)

缺点是大数据上难以计算 (computationally intractable for large dataset)

非参数机器学习算法：可以自由地从训练数据中学习任意形式目标函数的算法

寻求在构造目标函数过程中对训练数据作最好拟合，同时维持泛化到未知数据的能力

优势：可变性：可以拟合不同的目标函数形式

模型强大：对于目标函数不作假设或者作微小的假设

表现良好：对于预测表现可以非常好

局限性：数据更多：对于拟合目标函数需要更多的训练数据

速度慢：由于需要训练更多的参数，训练过程通常比较慢

过拟合：发生过拟合的风险更高，对于预测也比较难以解释

Machine Learning - P4

K最近邻 (K-nearest neighbour, KNN), 最简单的非参数分类方法 (non-parametric classifier)

核心思想是样本在特征空间中的 k 个最相邻的样本中大多数属于一个类别

则该样本也属于这个类别，且具有这个类别的样本的特性

石角定分类决策只依据最邻近的 k 个样本的类别来决定待分样本的所属类别

分类决策时只与极少量的相邻样本有关

主要依靠周围有限的邻近的样本，而非靠判别类域的方法来确定所属类别

适合于类域的交叉或重叠较多的待分样本

对于给定的输入向量 \vec{x} ，训练集 D，整数参数 K

令集合 $N_K(\vec{x}, D)$ 表示在训练集 D 中距离 \vec{x} 最近的 K 个样本

$I(E)$ 表示对于事件 E 的指示器变量

则输出 y 为一个分类变量 $y \in \{1, 2, \dots, C\}$

则有 $P(y=c | \vec{x}, D, K) = \frac{1}{K} \sum_{i \in N_K(\vec{x}, D)} I(y_i=c)$

应用最大后验概率估计 (MAP)

则 $\hat{y}(\vec{x}) = \operatorname{argmax}_c P(y=c | \vec{x}, D, K)$

KNN 是一种基于实例的学习 (instance-based learning)

或基于记忆的学习 (memory-based learning)

KNN 回归 (KNN regression)：输出为对象的属性值，通常为 K 个邻居的值的平均值

衡量量邻居的权重，可以使较近的邻居比较远的邻居的权重大

常见的方案是为每个邻居赋权重为 $1/d$ ，其中 d 是与邻居的距离

对于连续变量，通常使用欧几里得距离量 (Euclidean distance)

而对于离散变量，可以使用重叠度量 / 汉明距离 (Hamming distance)

K 值的选取可以通过启发式技术，最优的 K 值取决于数据

较大的 K 值可以减小噪声影响，但会使类别之间界限模糊

优点：简单，易于理解，易于实现，无需估计参数，无需训练

适合对稀有事件分类

特别适合于多分类问题

缺点：“多数表决”在分布偏斜时出现缺陷

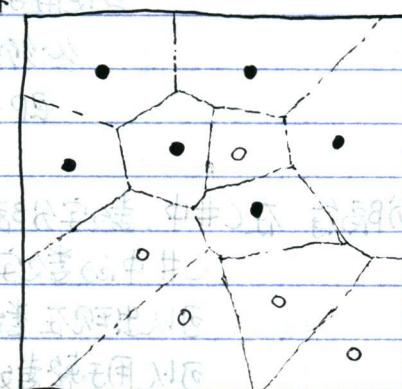
计算量较大，需要所有样本距离

可理解性差，无法给出如决策树的规则

沃罗诺伊图

沃罗诺伊 (Voronoi diagram)，当 K=1 时，则只取决于最近邻

(Voronoi tessellation)



Machine

Learning - P5

维数灾难 (curse of dimensionality), 又称维度的诅咒 (no curse of dimensionality)

用于描述当 (数学) 空间维度增加时, 分析和组织高维空间, 因为体积指数地增长而引发各种问题场景

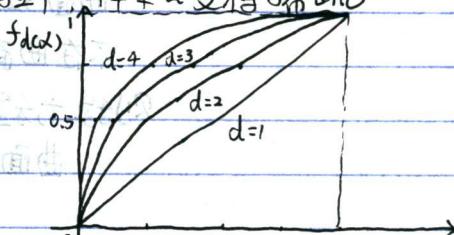
特性是从低维 (特征少) 转向高维 (特征多) 的过程中, 样本会变得稀疏

1. 样本数目不变, 但样本相互之间距离增大

2. 样本密度不变, 所需样本的数目指数级增长

假设训练集 D 的输入有 d 个特征

且每个特征的取值范围都分布在 $[0, 1]$



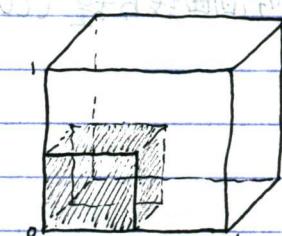
训练集 D 中的输入均是唯一的, 即互不相同

对于给定的特征空间的一个比例 $\alpha \in [0, 1]$

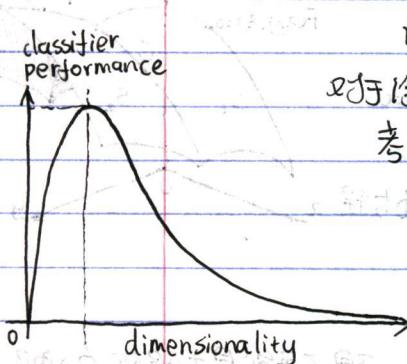
考虑覆盖特征空间的比例不低于 α 时

每个特征需要覆盖的平均比例 $f_d(\alpha)$

于是可知对于 $d \in \mathbb{N}^+$ 和 $\alpha \in [0, 1]$, 有 $f_d(\alpha) = \alpha^{1/d}$



即当 $\alpha > 0$ 时, $\lim_{d \rightarrow \infty} f_d(\alpha) = 1$

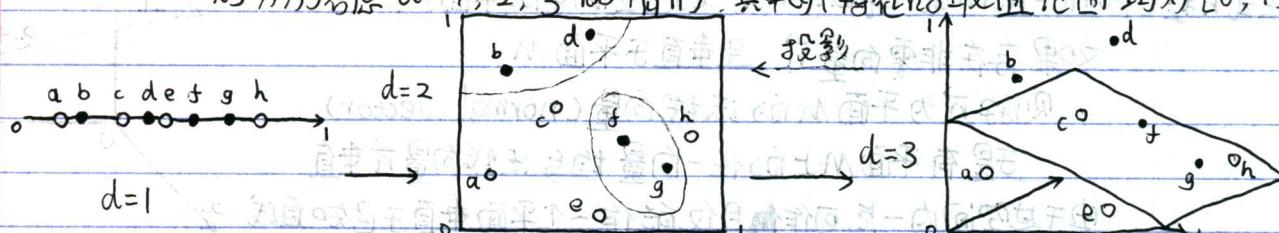


另一个影响是, 由于样本变得稀疏, 于是在低维不可分的样本在高维变得可分

即在合适的高维度下可以找到一个合适的可分的超平面

假设训练集 D 中有 8 个样本, 而输出分类变量 $y \in \{0, 1\}$

则分别考虑 $d=1, 2, 3$ 的情形, 其中每个特征的取值范围均为 $[0, 1]$



虽然随着维度增加, 越来越容易得到一个超平面来区分目标
但将高维度向低维度投影时会出现问题

过拟合 (overfitting), 当特征空间维度过高时

由于高维度的线性分类器, 相当于低维度的复杂非线性分类器

于是甚至会对错误/异常的数据进行分类学习

而正确数据无法覆盖整个特征空间,

从而导致训练集上表现良好但对新数据缺乏泛化能力

另外高维度用欧几里得距离来衡量样本的方法可能失效

高维度下任意两点间的距离会趋向收敛, 其最大最小距离会趋于相同

于是无法进行基于欧几里得距离的分类

Machine

Learning - P6

信息论

(information theory) 由克劳德·香农 (Claude Shannon) 发展

涉及信息的量化，存储，通信，用以找出信号处理与通信操作的基本限制，并扩展至其他领域
“数据的意义方面是无关紧要的”，而数据的性质和意义在信息内容方面并不重要
以概率分布 (probability distribution) 和不确定性 (uncertainty) 量化信息
并引入了比特 (bit) 的概念

自信息

(self-information), 也称信息量，是与概率空间中的单一事件或离散随机变量值相关的信息量的

以信息的单位 bit / nat (奈特) / hart 表示，取决于计算中使用的对数的底

对于给定的随机变量 X 以及概率密度函数 $p(x)$, 其中 $x \in$ 样本空间 S_x

则对于给定的结果 $x \in S_x$, 记 $I_X(x)$ 为 x 的自信息.

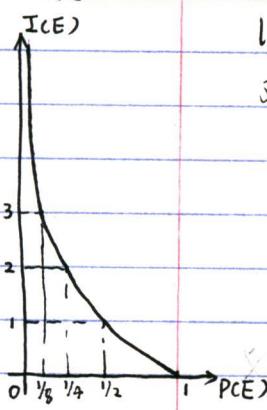
$$I_X(x) = -\log [P(x)] = \log \left(\frac{1}{p(x)} \right)$$

更一般地对于事件 E , 其发生概率为 P

$$\text{则 } I(E) = -\log [\Pr\{E\}] = -\log(P) = \log \left(\frac{1}{P} \right)$$

自信息的单位由进行对数运算时的底决定

2: bit (比特), e: nat (奈特) \log_{10} : hartley/hart (哈特利).



对于概率的

反单调性 (antitonicity for probability). 对于给定的概率空间 (probability space)

观测到更罕见的事相比更常见的事，会提供更多信息

measurement of rarer event yield more information content than more common event

于是自信息表现出观测下的事件概率的反单调性

antitonic in probability for event under observation

独立事件的可加性 (additivity of independent events), 也称 sigma additivity

对于相互独立的随机变量 X 和 Y , 分别有概率密度函数 $p(x)$ 和 $p(y)$

则其联合概率密度函数 $p(x,y) = \Pr\{X=x\} \Pr\{Y=y\} = p(x)p(y)$

是对于给定的 $(x,y) \in S_x \times S_y$, 记 $I_{X,Y}(x,y)$ 为 (x,y) 的自信息

则 $I_{X,Y}(x,y) = -\log [P(x,y)] = -\log [p(x)p(y)]$

$$= -\log [p(x)] - \log [p(y)] = I_X(x) + I_Y(y)$$

这种度量也称为 surprisal, 相对竟的性质是似然 (likelihood)

独立事件的联合对数似然 (log-likelihood) 是各自对数似然的和

log-likelihood 也可以描述为 support 或 negative surprisal

degree to which event support given model

model supported by event to the extent that event unsurprising, given the model

Machine

Learning - P7

自信息

根据自信息的定义，对于信息的生成者和接收者，以及信息的“消息载体”。

仅当接收者未提前知道信息时，才有信息从生成者传递到接收者。

information transformed from originating entity to receiving entity

only when receiver has not know the information a priori

当消息内容已被提前确定地知道，即概率为1，则消息中没有传达任何信息。

when content of message known a priori with certainty, with probability of 1

no actual information conveyed in the message

仅当消息内容中的增量信息对于接收者并非完全确定的，消息才传递了信息。

only when advance knowledge of content of message < 100% certain by receiver
does the message actually convey information

信息熵 (information entropy)，也称香农熵 (Shannon entropy)，用于表示信息量的期望。

对于给定的随机变量 X 以及概率密度函数 $P(x_i)$ ，样本空间为 S_X

则记 $H(X)$ 为随机变量 X 的信息熵

于是有 $H(X) = E[I(X)] = \sum_{x_i \in S_X} P(x_i) I_x(x_i) = -\sum_{x_i \in S_X} P(x_i) \log P(x_i)$

特别地对于成功率 p 的伯努利试验，称为 binary entropy function

有 $H_b(p) = -p \log p - (1-p) \log(1-p)$

于是信息熵描述了一个分布的信息量，或者编码的平均长度。

在香农的信源编码理论 (Shannon's source coding theorem) 中

信息熵提供了压缩率的下限。

即当使用少于信息熵的信息量做编码时，则一定有信息的损失。

另外对于样本空间为 $\{x_1, x_2, \dots, x_n\}$ 的离散随机变量 X 和概率分布函数 $P(x_i)$

有 $H(X) = -\sum_{i=1}^n P(x_i) \log P(x_i)$

联合熵 (joint entropy)，对于两个离散随机变量 X, Y ，有随机变量为有序对 (X, Y)

则 (X, Y) 的信息熵称为随机变量 X, Y 的联合熵，记为 $H(X, Y)$

(X, Y) 服从概率密度函数 $P(x, y) = \Pr\{X=x \wedge Y=y\}$

则有 $H(X, Y) = E[I_{X,Y}(x, y)] = -\sum_{(x,y)} P(x, y) \log P(x, y)$

条件熵 (conditional entropy)，对于两个离散随机变量 X, Y ，记 $H(X|Y)$ 为 X 在给定 Y 的条件熵。

则有 $H(X|Y) = E_Y[H(X|y)] = -\sum_y P(y) \sum_x P(x|y) \log P(x|y)$

$= -\sum_{x,y} P(x, y) \log P(x|y)$

即 $H(X|Y)$ 为给定 $Y=y$ 的随机变量 X 的信息熵的期望。

Machine

Learning - P8

链式法则

对于两个离散随机变量 X, Y , 有联合熵 $H(X, Y)$, $H(Y, X)$ 和条件熵 $H(X|Y)$, $H(Y|X)$

由于 $P(x|y) = P(x, y) / P(y)$, 且 $P(x, y) = P(y, x)$

于是有 $H(X|Y) = E_{(x,y)} P(x, y) \log P(x, y)$

$$\begin{aligned} H(X|Y) &= \sum_{(x,y)} P(x, y) \log [P(x|y) P(y)] \\ &= \sum_{(x,y)} P(x, y) \log P(x|y) + \sum_y [\sum_x P(x, y)] \log P(y) \\ &= H(X|Y) + H(Y) \end{aligned}$$

同理地有 $H(Y|X) = H(Y|X) + H(X)$

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

$H(X, Y)$	
$H(X)$	$H(Y X)$
$H(X Y)$	$H(Y)$
$H(Y, X)$	

← 信息量 →

扩展至对于 n 个随机变量 $X_1, X_2, X_3, \dots, X_n$, 有联合熵 $H(X_1, X_2, \dots, X_n)$

则参考对 n 个事件 E_1, E_2, \dots, E_n 的交的频率的乘法法则

即有 $\Pr\{E_1, E_2, \dots, E_n\} = \Pr\{E_1\} \Pr\{E_2|E_1\} \Pr\{E_3|E_1, E_2\} \dots \Pr\{E_n|E_1, E_2, \dots, E_{n-1}\}$

$$= \Pr\{E_1\} \cdot \prod_{i=2}^n \Pr\{E_i | \cap_{j=1}^{i-1} E_j\}$$

对根据联合熵与条件熵的关系 $H(X, Y) = H(Y|X) + H(X)$

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &= H(X_1) + H(X_2 | X_1) + H(X_3 | X_1, X_2) + \dots + H(X_n | X_1, \dots, X_{n-1}) \\ &= H(X_1) + \sum_{i=2}^n H(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

$H(X_1)$	$H(X_2 X_1)$	$H(X_3 X_1, X_2)$	\dots	$H(X_n X_1, \dots, X_{n-1})$
顺序地接收 X_1, X_2, \dots, X_n				

于是收到一个序列 $\{X_1, X_2, \dots, X_n\}$ 的信息量可以看作

顺序地接收 X_1, X_2, \dots, X_n 时的信息增量的和

称为信息熵的链式法则

互信息

(mutual information), 或称信息增益 (transformation)

对于离散变量 X, Y , 记 $I(X; Y)$ 为 X, Y 的互信息

$$\text{则有 } I(X; Y) = E[S I(x, y)] = \sum_{(x,y)} P(x, y) \log \frac{P(x, y)}{P(x) P(y)}$$

其中 $S I(x, y)$ 为 (Specific mutual Information), 是 pointwise mutual information

描述了联合分布中两个信息的纠缠程度 / 相互影响部分的信息量

$$\text{且有 } I(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) = H(Y) - H(Y|X) = I(Y; X)$$

当 X, Y 相互独立时, 有 $H(X, Y) = H(X) + H(Y)$, 即 $I(X; Y) = I(Y; X) = 0$

$I(X; Y) \leq \min(H(X), H(Y))$, 当且仅当存在一一对应 $g: X \rightarrow Y$ 取得等号

注意此时 X, Y 完全相关, $H(X) = I(X; Y) = I(Y; X) = H(Y) = H(X, Y)$

$H(X)$	$H(Y X)$
$H(X Y)$	$I(X; Y)$
$H(X Y)$	$H(Y)$
$H(X, Y)$	

← 信息量 →

Machine

Learning - P9

对于离散随机变量 X , 样本空间为 $\{x_1, x_2, \dots, x_n\}$, 其中 $n \in \mathbb{Z}^+$, 有概率密度函数 $p(x_i)$
则有信息熵 $0 \leq H(X) \leq \log n$

证明过程有, 通过拉格朗日乘子法证明

首先有 $p(x_1) + p(x_2) + \dots + p(x_n) = 1$

且对任意 $1 \leq i \leq n$, 有 $0 < p(x_i) \leq 1$

$$\text{则 } -p(x_i) \log p(x_i) \geq 0$$

$$\text{于是可知 } H(X) = \sum_{i=1}^n -p(x_i) \log p(x_i) \geq 0$$

$$\text{再取目标函数 } f[p(x_1), p(x_2), \dots, p(x_n)] = \sum_{i=1}^n -p(x_i) \log p(x_i)$$

$$\text{约束条件 } g[p(x_1), p(x_2), \dots, p(x_n), c] = [p(x_1) + p(x_2) + \dots + p(x_n) - 1] = 0$$

$$\text{则令拉格朗日函数 } L[p(x_1), p(x_2), \dots, p(x_n), \lambda] = f[p(x_1), \dots, p(x_n)] + \lambda g[p(x_1), \dots, p(x_n)] \\ = \sum_{i=1}^n -p(x_i) \log p(x_i) + \lambda [p(x_1) + p(x_2) + \dots + p(x_n) - 1]$$

分别求 $L[p(x_1), \dots, p(x_n), \lambda]$ 对 $p(x_1), \dots, p(x_n), \lambda$ 的偏导数, 并令偏导数为 0

$$\text{则 } \frac{\partial L[p(x_1), \dots, p(x_n), \lambda]}{\partial p(x_i)} =$$

$$= \frac{\partial [-p(x_i) \log p(x_i) + \lambda p(x_i)]}{\partial p(x_i)}$$

$$= -\log p(x_i) - p(x_i) \cdot \frac{\log e}{p(x_i)} + \lambda$$

$$\text{于是有 } \begin{cases} \lambda - \log [e p(x_i)] = 0, \lambda - \log [e p(x_2)] = 0, \dots, \lambda - \log [e p(x_n)] = 0 \\ p(x_1) + p(x_2) + \dots + p(x_n) - 1 = 0 \end{cases}$$

$$\text{则有 } p(x_1) = p(x_2) = \dots = p(x_n) = \frac{1}{n}$$

于是有时函数 $f[p(x_1), \dots, p(x_n)]$ 的极值为

$$f[\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}] = \sum_{i=1}^n -\frac{1}{n} \log \frac{1}{n} = -\log \frac{1}{n} = \log n$$

即可知 $0 \leq H(X) \leq \log n$

KL 散度 (Kullback-Leibler divergence), 又称相对熵 (relative entropy)

对于离散变量 X , 有 $p(x)$ 和 $q(x)$ 两种不同的概率分布的概率密度函数

$$\text{则记 } D_{KL}(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p(\log \frac{p(x)}{q(x)})$$

即 KL 散度是 $p(x)$ 与 $q(x)$ 之间对数差在 $p(x)$ 上的期望值

相对熵用于衡量两个概率分布之间的差异

当且仅当 $p(x)$ 与 $q(x)$ 分布相同时, 则有 $D_{KL}(p \parallel q) = 0$

相对熵不具有对称性, 即 $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$

$$\text{非负性: } D_{KL}(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_x p(x) \log \frac{q(x)}{p(x)} = -E_p(\log \frac{q(x)}{p(x)})$$

$$\geq -\log E_p(\frac{q(x)}{p(x)}) = -\log \sum_x p(x) \cdot \frac{q(x)}{p(x)}$$

$$(=-\log \sum_x q(x)) = -\log 1 = 0$$

Machine

Learning - P10

交叉熵

(cross entropy), 对于离散变量 X , 样本空间为 S_X

有两种不同的概率分布, 概率密度函数分别为 $p(x)$, $q(x)$

则有 KL 散度 $D_{KL}(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

于是有 $D_{KL}(P||Q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x)$

$= [-\sum_x p(x) \log q(x)] - H_p(X)$

则记 $H(p, q) = -\sum_x p(x) \log q(x)$

称为分布 P, Q 的交叉熵, 描述了两个分布的相似程度

于是有 $H(p, q) = H_p(X) + D_{KL}(P||Q)$



$H_p(X)$	$D_{KL}(P Q)$
$H(p, q)$	

→ $p(x)$ 的信息量 \times $p(x), q(x)$ 的 KL 散度

与 KL 散度类似, 交叉熵也不具有对称性

即 $H(p, q) \neq H(q, p)$

另外交叉熵描述两个分布之间的相互关系, 对分布自身求交叉熵时等价于信息熵

即当 $p(x)$ 与 $q(x)$ 是完全相同的分布时,

$H(p, p) = H_p(X)$, 即此时 KL 散度 $D_{KL}(P||P) = 0$

特别地有, 当分布 $p(x)$ 表示一个常量时,

即存在 $1 \leq i \leq n$, 有 $p(x_i) = 1$ 且 $\forall j \in \{1, \dots, n\} \setminus \{i\}, p(x_j) = 0$

则此时 $H_p(X) = 0$, 于是有 $D_{KL}(P||Q) = H(p, q)$

于是有“不严谨的”概念描述:

信息熵: 描述一个分布 $p(x)$ 的信息量期望, 即分布 $p(x)$ 包含的信息

KL 散度: 描述从分布 $p(x)$ 的角度看, 分布 $q(x)$ 有多大不同

交叉熵: 描述从分布 $p(x)$ 的角度看, 如何描述分布 $q(x)$

而 KL 散度可以用于度量代价

对于机器学习, 有真实分布 P_{real} , 训练集的分布 $P_{training}$, 模型学习得到的分布 P_{model}

最终目的是模型学到的分布与真实一致, 即 $P_{model} \approx P_{real}$

而真实分布不可知, 于是假设训练集分布与真实一致, 即 $P_{training} \approx P_{real}$

于是目的转为希望模型分布与训练集分布一致, 即 $P_{model} \approx P_{training} \approx P_{real}$

则需要最小化 P_{model} 和 $P_{training}$ 的 KL 散度 $D_{KL}(P_{training} || P_{model})$

由于 $P_{training}$ 是已知的, 即 $H_{P_{training}}(X) = 0$

于是求 $D_{KL}(P_{training} || P_{model})$ 等价于求 $H(P_{training}, P_{model})$

则当交叉熵 $H(P_{training}, P_{model}) = 0$ 时, 认为学到了“最好的模型”

但是完美学到训练集通常意味着过拟合, 因为训练集不等于真实数据

通常还需假设存在一个正态分布的误差, 是模型泛化误差下限

Machine

Learning - P11

泊松分布的 极大似然估计 对于包含 n 个样本的训练集 $D = \{x_1, x_2, \dots, x_n\}$ 假设 $x_i \sim \text{Poisson}(\lambda)$, 且 x_1, x_2, \dots, x_n 服从独立同分布, 其中 $i=1, 2, \dots, n$

则估计泊松分布的参数 λ

首先有概率密度函数 $P(x_i; \lambda) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}, i=1, 2, \dots, n, x_i \in \mathbb{N}$

于是取对数似然函数 $L(\lambda) = \log \Pr\{D|\lambda\}$

$$\begin{aligned} L(\lambda) &= \log \Pr\{D|\lambda\} = \log [\prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}] \\ &= \sum_{i=1}^n \log [e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}] \\ &= \sum_{i=1}^n [\log e^{-\lambda} + \log \lambda^{x_i} - \log x_i!] \\ &= \sum_{i=1}^n -\lambda + \sum_{i=1}^n x_i \log \lambda - \sum_{i=1}^n \log x_i! \end{aligned}$$

取对参数 λ 的 $L(\lambda)$ 的导数, 并取导数为 0

$$\begin{aligned} \text{于是 } \frac{dL(\lambda)}{d\lambda} &= (\sum_{i=1}^n -\lambda + \sum_{i=1}^n x_i \log \lambda - \sum_{i=1}^n \log x_i!)' \\ &= -n + \frac{1}{\lambda} \sum_{i=1}^n x_i = 0 \end{aligned}$$

$$\text{即 } \lambda = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

于是可知对参数 λ 的估计为训练集 $D = \{x_1, \dots, x_n\}$ 的期望 \bar{x}

指数分布的 极大似然估计 对于包含 n 个样本的训练集 $D = \{x_1, x_2, \dots, x_n\}$

假设 $x_i \sim \text{Exponential}(\theta), i=1, 2, \dots, n$, 且 x_1, x_2, \dots, x_n 服从独立同分布

则估计指数分布的参数 θ

首先有概率密度函数 $P(x_i; \theta) = \begin{cases} \theta e^{-\theta x_i}, & x_i \geq 0, i=1, 2, \dots, n \\ 0, & x_i < 0 \end{cases}$

于是取对数似然函数 $L(\theta) = \log \Pr\{D|\theta\}$

$$\begin{aligned} \text{则 } L(\theta) &= \log \Pr\{D|\theta\} = \log [\prod_{i=1}^n \theta e^{-\theta x_i}] \\ &= \sum_{i=1}^n \log [\theta e^{-\theta x_i}] \\ &= \sum_{i=1}^n \log \theta + \sum_{i=1}^n (-\theta x_i) \\ &= n \log \theta - \theta \sum_{i=1}^n x_i \end{aligned}$$

取对参数 θ 的 $L(\theta)$ 的导数, 并取导数为 0

$$\text{于是 } \frac{dL(\theta)}{d\theta} = (n \log \theta - \theta \sum_{i=1}^n x_i)' = 0$$

$$\frac{1}{\theta} n - \sum_{i=1}^n x_i = 0$$

$$\theta = \frac{n}{\sum_{i=1}^n x_i}$$

$$\text{即有 } \theta = \frac{1}{\bar{x}}$$

于是可知对参数 θ 的估计为训练集 $D = \{x_1, x_2, \dots, x_n\}$ 期望的倒数 $\frac{1}{\bar{x}}$

注意此处的指数运算 \log 均为以 e 为底, 即 \ln