PRML

1 绪论

泛化能力：正确分类与训练集不同的新样本的能力.

强化学习：关注在给定的条件下，找到合适的动作，使得奖励达到最大值.

1.1 例子: 多项式拟合

用多项式函数：$y(x, \boldsymbol{w}) = \sum_{i=0}^{M} w_i x_i$.

- 预测值与目标值的平方和误差：

  损失函数求导后可以得到闭式解.

  多项式的次数M是超参数，模型选择问题. 过拟合：M较大时损失函数为零，但是曲线剧烈震荡.

- 均方根误差：

  预测值与目标值的平方和 $\div$ 样本个数，再开根号.

  以相同的基础对比不同大小的数据集.

随着M的增大，系数通常会增大(暗示曲线会震荡)，但是多项式函数可以精确拟合训练集. 过分地拟合了随机噪声.

数据集规模增加，过拟合问题变得不那么严重. 数据集规模越大，我们能够用来拟合数据的模型就越复杂(即越灵活).

正则化:

减少过拟合.

- 岭回归：平方和误差 $+$ 正则化项 $\lambda\|\boldsymbol{w}\|^2$.

1.2 概率论

- 关于一个条件分布的条件期望：

$$\mathbb{E}_x[f \mid y] = \sum_x p(x \mid y) f(x)$$

1.2.3 贝叶斯概率

- 先验：$p(\boldsymbol{w})$，在观察数据之前，我们有一些关于参数(比如多项式曲线例子中的$\boldsymbol{w}$)的假设．
- 类条件概率/似然：$p(\mathcal{D} \mid \boldsymbol{w})$，表达在不同的参数$\boldsymbol{w}$下，观测数据出现的概率．

  似然函数不是$\boldsymbol{w}$的概率分布，关于$\boldsymbol{w}$的积分不一定等于1．

$$p(\boldsymbol{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{w}) p(\boldsymbol{w})}{p(\mathcal{D})}$$

- 频率学家：最大似然估计．

  $\boldsymbol{w}$ 是一个固定的参数，由某种形式(比如数据集$\mathcal{D}$的概率分布)的"估计"来确定．
- 贝叶斯：包含先验概率．实际观测到数据集$\mathcal{D}$，参数的不确定性用$\boldsymbol{w}$的概率分布来表达．该方法的缺点：选择先验概率通常是选方便的而不是反映出先验知识的．

1.2.4 高斯分布

高维高斯：

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

$$\mathcal{N}(\vec{x}|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\cdot\exp\left\{-\frac{1}{2\sigma^2}(\vec{x}-\mu)^2\right\}$$

$$\mathcal{L}=\ln\prod_i^N\mathcal{N}(x_i|\mu,\sigma^2)$$

$$=-\frac{N}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^N(x_i-\mu)^2$$

$$\frac{\partial\mathcal{L}}{\partial\mu}=0\Rightarrow\mu=\frac{1}{N}\sum_i^N x_i\ ;\ \frac{\partial\mathcal{L}}{\partial\sigma^2}=0\Rightarrow\sigma^2=\frac{1}{N}\sum_{i=1}^N(x_i-\mu)^2$$

$$-\frac{N}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_i^N=0$$

精度参数：$\beta=\frac{1}{\sigma^2}$

(MLE) $\mathcal{L}=\ln p(y_i|x,w,\beta)=\ln\left(\prod_i^N\mathcal{N}(y_i|f(x_i,w),\frac{1}{\beta})\right)$

真值

$$=-\frac{N}{2}\ln(2\pi\cdot\frac{1}{\beta})-\frac{\beta}{2}\sum_i^N(f(x_i,w)-y_i)^2$$

☆MLE可以得到似然：$p(y|x,w,\beta)=\mathcal{N}(y|f(x,w_{MLE}),\beta_{MLE}^{-1})$

引入系数w的先验：$p(w|\alpha)$，其中α是分布的精度.

贝叶斯：

$$p(w|x,y,\alpha,\beta)\propto p(y|x,w,\beta)\cdot p(w|\alpha)$$

真值

☆给w的先验是高斯：　协方差

$$p(w|\alpha)=\mathcal{N}(w|0,\frac{1}{\alpha}E)$$

$$=\left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}}\exp\left\{-\frac{\alpha}{2}w^Tw\right\}$$

所以最大似然估计的均值可正确(是无偏估计)，但是方差被系统性地低估了(不是无偏估计).

μ_{MLE} 与 σ^2_{MLE} 无关.

最大似然的偏移问题是我们在多项式曲线拟合问题中遇到的过拟合问题的核心：

1.2.5 重新考察曲线拟合问题

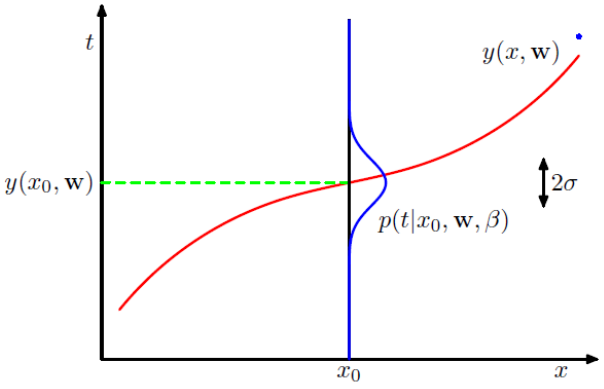
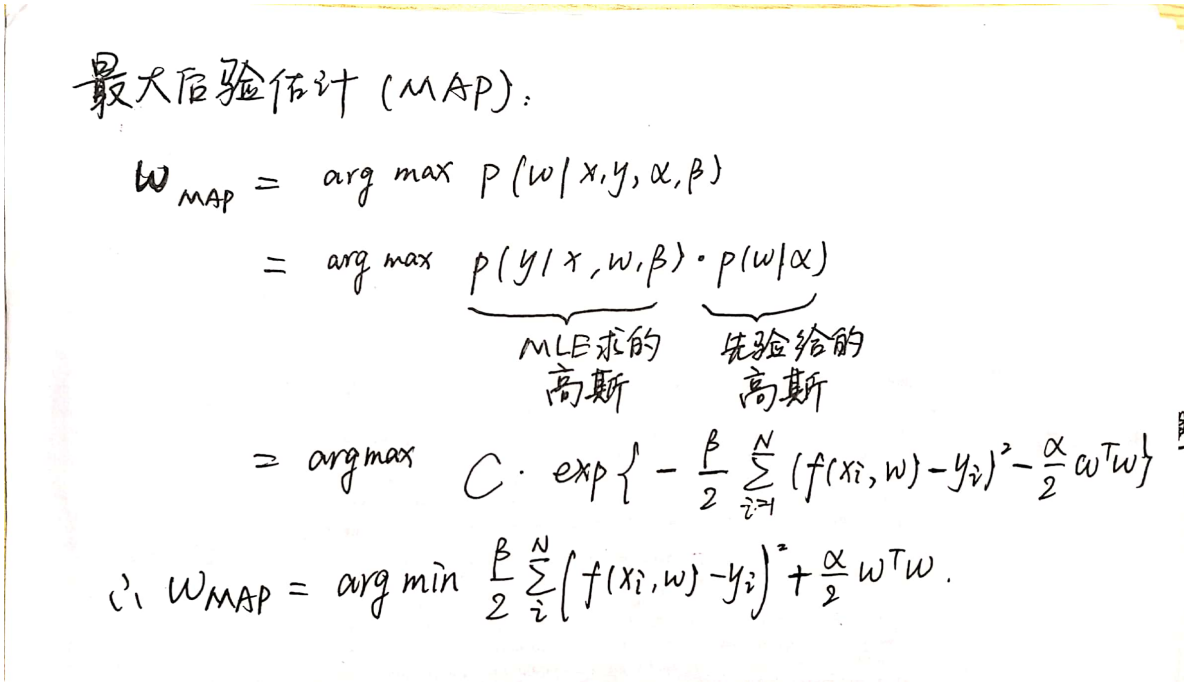


Figure 1.16 Schematic illustration of a Gaussian conditional distribution for t given x given by (1.60), in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the precision is given by the parameter β, which is related to the variance by $\beta^{-1} = \sigma^2$.

很牛！



所以：

- 在$\boldsymbol{w}$的先验是高斯，就是上面图上那个曲线y轴方向的高斯．
- 并且MLE作用于似然函数．
- **最后MAP求解后的结果 $\Rightarrow$ 岭回归．**这就是贝叶斯视角下的岭回归．

1.2.6 贝叶斯曲线拟合

更纯的贝叶斯：

贝叶斯曲线拟合

参数后验：$p(w|x,y)$，预测概率：$p(y'|x')$

二、预测分布：$p(y'|x',x,y) = \mathcal{N}(y|m(x), s^2(x))$

（参数后验的 x）

其中：$m(x) = \beta \phi(x)^T S \sum_i^N \phi(x_i) y_i$

$s^2(x) = \beta^{-1} + \phi(x)^T S \phi(x)$

$S = \alpha E + \beta \sum_i^N \phi(x_i)\phi(x_i)^T$

1.3 模型选择

交叉验证：平均分成几份，留一份做验证集，其他的是训练集，如此交替．

信息准则尝试修正最大似然的偏差，增加惩罚项：

- Akaike information criterion(AIC):

  $$\ln p(\mathcal{D}|w_{MLE}) - M$$

  第一项是对数似然，M 是模型中可调节参数的数量．

  最大化上式．

1.4 维度灾难

- 例子：将输入空间划分成一个个单元格，预测输入的类别就是它所在单元格中其他数据最多的类别．

  上例中单元格的数量随着空间的维数指数增大．为了保证单元格不空，需要指数量级的训练数据．
- 多项式曲线拟合例子：系数数量的增长速度类似于 D^M．
- 考虑 $r = 1-\epsilon$ 到 $r = 1$ 之间的体积 占超球总体积的百分比：

  $$V_D(r) = C_D \cdot r^D$$

  $$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

  发现对于较大的 D，即使是很小的ϵ，占比也趋近于1，即大部分体积都聚集在表面附近的超球壳上．

但是仍有应用于高维空间的有效技术：

1. 真实数据经常被限制在有着较低的有效维度的空间区域中．
2. 真实数据通常比较光滑．大多数情况输入数据微小改变，目标值改变也很小．

1.5 决策论

先验：$p(\mathcal{C}_k)$．

后验：$p(\mathcal{C}_k|x)$.

输入与真值(真实类别)的不确定性：$p(\boldsymbol{x},\mathcal{C}_k)$.

1.5.1 最小化错误分类率

(两类)错误分类的概率，(多类)正确分类的概率：

$$\begin{aligned} p(\text{ mistake }) &= p\left(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2\right)+p\left(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\ &= \int_{\mathcal{R}_1} p\left(\boldsymbol{x}, \mathcal{C}_2\right) \mathrm{d} \boldsymbol{x}+\int_{\mathcal{R}_2} p\left(\boldsymbol{x}, \mathcal{C}_1\right) \mathrm{d} \boldsymbol{x} \\ p(\text{ correct }) &= \sum_k p\left(\boldsymbol{x} \in \mathcal{R}_k, \mathcal{C}_k\right)=\sum_{k=1}^K \int_{\mathcal{R}_k} p(\boldsymbol{x}, \mathcal{C}_k) d \boldsymbol{x} \end{aligned}$$

其中$\mathcal{R}$是根据$\boldsymbol{x}$的类别划分的决策区域.

图例(很牛!)：

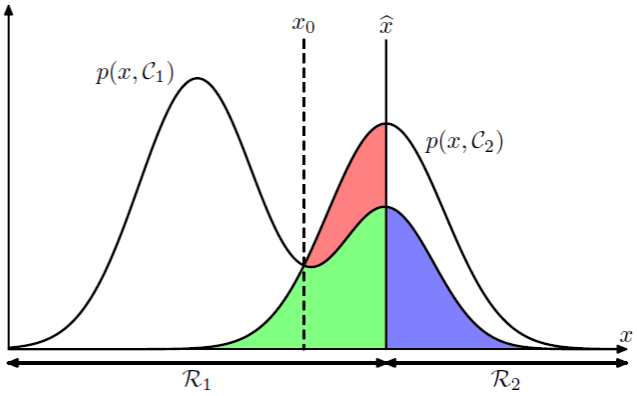


Figure 1.24 Schematic illustration of the joint probabilities $p(x,\mathcal{C}_k)$ for each of two classes plotted against x, together with the decision boundary $x=\widehat{x}$. Values of $x \geqslant \widehat{x}$ are classified as class $\mathcal{C}_2$ and hence belong to decision region $\mathcal{R}_2$, whereas points $x<\widehat{x}$ are classified as $\mathcal{C}_1$ and belong to $\mathcal{R}_1$. Errors arise from the blue, green, and red regions, so that for $x<\widehat{x}$ the errors are due to points from class $\mathcal{C}_2$ being misclassified as $\mathcal{C}_1$ (represented by the sum of the red and green regions), and conversely for points in the region $x \geqslant \widehat{x}$ the errors are due to points from class $\mathcal{C}_1$ being misclassified as $\mathcal{C}_2$ (represented by the blue region). As we vary the location $\widehat{x}$ of the decision boundary, the combined areas of the blue and green regions remains constant, whereas the size of the red region varies. The optimal choice for $\widehat{x}$ is where the curves for $p(x,\mathcal{C}_1)$ and $p(x,\mathcal{C}_2)$ cross, corresponding to $\widehat{x}=x_0$, because in this case the red region disappears. This is equivalent to the minimum misclassification rate decision rule, which assigns each value of x to the class having the higher posterior probability $p(\mathcal{C}_k|x)$.

- 上图：

  小于$\hat{x}$，被分类为$\mathcal{C}_1$.

  横轴标的是预测类别的决策区域.

  - 把 $\mathcal{C}_1$ 分到 $\mathcal{C}_2$：蓝色区域，就是$p(x,\mathcal{C}_1)$在决策区域 $\mathcal{R}_2$ 的概率和.
  - 把 $\mathcal{C}_2$ 分到 $\mathcal{C}_1$：红色区域加绿色区域，同理这就是 $p(x,\mathcal{C}_2)$ 在决策区域 $\mathcal{R}_1$ 的概率和.

  预测时候就是改变 $\hat{x}$，注意到此时绿色和蓝色区域总和是常数. 红色区域面积在改变. 最优时候就是 $\hat{x}=x_0$. 等价于最小化错误分类率的决策规则.

1.5.2 最小化期望损失

损失(代价)矩阵 L：分类错误代价.

期望损失：

$$\mathbb{E}[L]=\sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} \cdot p\left(\boldsymbol{x}, \mathcal{C}_k\right) \mathrm{d} \boldsymbol{x}$$

我们的目标是划分到最优的 $\mathcal{R}_j$.

用贝叶斯定理转换为后验，就是西瓜书第七章前面的 最小化条件风险(风险就是期望损失)，对于一个$\boldsymbol{x}$，它可以被分到取值最小的第j类：

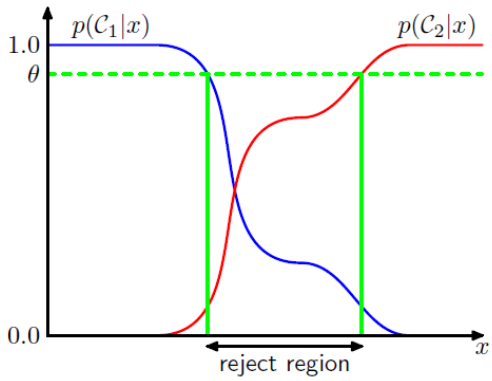
$$R\left(\mathcal{C}_j \mid \boldsymbol{x}\right)=\sum_k L_{kj} \cdot p\left(\mathcal{C}_k \mid \boldsymbol{x}\right)$$

1.5.3 拒绝选项

对于难以分类的情况，拒绝分类，交给人类.

后验概率小于某个阈值 θ，则拒绝分类：

Figure 1.26 Illustration of the reject option. Inputs x such that the larger of the two posterior probabilities is less than or equal to some threshold θ will be rejected.



上图中间那块，如果后验中较大的那个还是小于θ，就会落入拒绝区域.

1.5.4 推断和决策

- 推断阶段：使用训练数据学习 $p(\mathcal{C}_k|\boldsymbol{x})$ 的模型.
- 决策阶段：使用后验概率进行最优分类.

1.5.5 回归问题的损失函数

期望损失：

$$\mathbb{E}[L]=\int\int L(y, f(\boldsymbol{x})) \cdot p(\boldsymbol{x}, t)\, d\boldsymbol{x}\, dt$$

其中 L 可以是平方损失.

变分法：（其中 $L=(f(x)-y)^2$，$f(x)$预测，y真值）

$$\frac{\delta \mathbb{E}[L]}{\delta f(x)} = 2\int (f(x)-y)\cdot p(x|y)\,dy = 0$$

$$\Rightarrow f(x)\int p(x|y)dy = \int y\cdot p(x|y)dy$$

$$\therefore f(x) = \frac{\int y\cdot p(x|y)\,dy}{p(x)} = \int y\cdot p(y|x)\,dy = \mathbb{E}_y[y|x]$$

在x条件下 y的条件期望

最优解是条件期望.

最优解是条件期望.

$$(f(x)-y)^2 = (f(x)-\mathbb{E}[y|x]+\mathbb{E}[y|x]-y)^2$$

$$= (f(x)-\mathbb{E}[y|x])^2 + (\mathbb{E}[y|x]-y)^2 + 2(f(x)-\mathbb{E}[y|x])(\mathbb{E}[y|x]-y)$$

对 x 积分：

$$\mathbb{E}[L] = \mathbb{E}[(f(x)-y)^2]$$

$$= \int (f(x)-\mathbb{E}[y|x])^2 p(x)\,dx + \int (\mathbb{E}[y|x]-y)^2 p(x)\,dx$$

$f(x)$ 仅出现在第一项中

当 $f(x)=\mathbb{E}[y|x]$ 时，最优解，与上述条件期望相符.

第二项：y 分布的类似方差的，可以看成数据噪声.

交叉项为零（对y积分）

$$\iint (\mathbb{E}[y|x]-y)\,p(x,y)\,dy\,dx \qquad \mathbb{E}(y|x) = \int y\frac{p(x,y)}{p(x)}dy$$

$$= \int \mathbb{E}(y|x)p(x)\,dx - \int \mathbb{E}(y|x)p(x)\,dx$$

$$= 0$$

回归问题也有三种解决方式：

- 推断 联合概率密度 $p(\boldsymbol{x},y)$，计算条件概率密度 $p(y|\boldsymbol{x})$，最后 $\int y\cdot p(y|\boldsymbol{x})\,dy$，求条件期望.
- 推断条件概率密度 $p(y|\boldsymbol{x})$，其他与上一致.
- 直接训练一个回归函数 $f(x)$

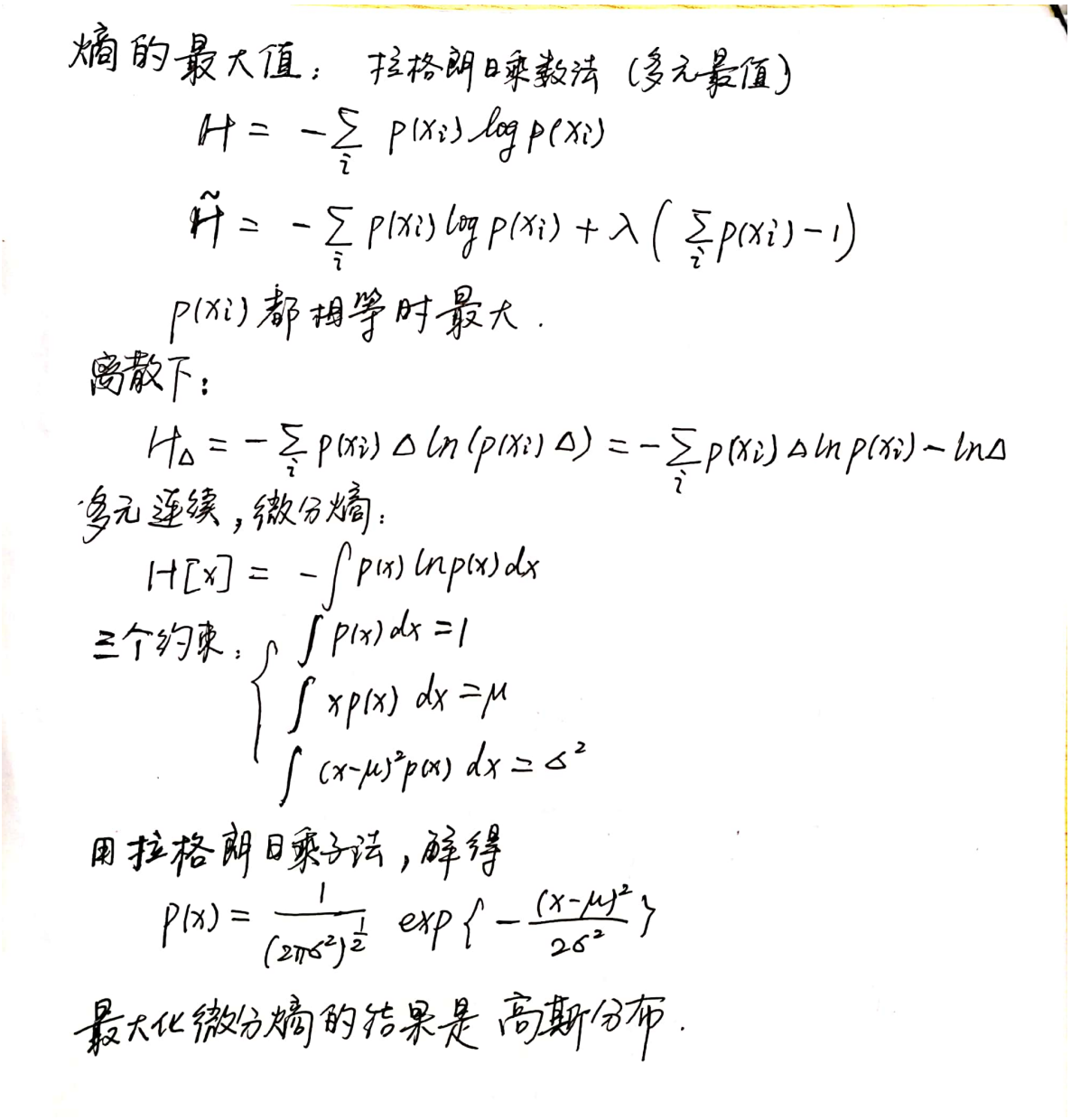
其他损失函数的期望，闵可夫斯基损失函数：

$$\mathbb{E}[L_q] = \iint |y(\mathbf{x})-t|^q \cdot p(\mathbf{x},t)\mathrm{d}\mathbf{x}\mathrm{d}t$$

其中 $q=2$ 时就是平方损失.

1.6 信息论

信息熵.



1.6.1 相对熵和互信息

- KL散度分布之间的相对熵，就是两个相减：

$$\begin{aligned}\mathrm{KL}(p\,\|\,q)&=-\int p(\boldsymbol{x})\ln q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\left(-\int p(\boldsymbol{x})\ln p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\\&=-\int p(\boldsymbol{x})\ln\left\{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right\}\mathrm{d}\boldsymbol{x}\end{aligned}$$

  KL散度不满足对称性.

- 琴生(Jensen)不等式：

  用琴生不等式和 $-ln(x)$ 是凸函数，证明KL散度非负：

$$\mathrm{KL}(p\|q)=-\int p(\boldsymbol{x})\ln\left\{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right\}\mathrm{d}\boldsymbol{x}\geq-\ln\int q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0$$

  放到ln里面就可.

数据由某未知分布 $p(x)$ 生成

我们用 $q(x|\theta)$ 来近似这个分布 $(p(x))$

最小化 KL 散度：$KL(p\|q) \simeq \frac{1}{N}\sum_{i}^{N}\left(-\ln q(x_i|\theta)+\ln p(x_i)\right)$

互信息：（联合概率分布与边缘概率分布之间是否"接近"相互独立）

$$I[x,y] = KL\left(p(x,y)\,\|\,p(x)p(y)\right)$$

$$= -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)dx\,dy.$$

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x].$$

1.7 练习