1 Probability and Bayes Estimation

1. The Binary Symmetric Channel (BSC) can be represented by the transition matrix:

$$P(Y|X) = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}.$$

Assuming the marginal probability $p(X=0) = 1-\beta$ and $p(X=1) = \beta$, compute the mutual information $I(X;Y)$.

Solution

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

由转移矩阵,

$$P(Y=0) = (1-\beta)(1-\epsilon) + \beta\epsilon, \quad P(Y=1) = (1-\beta)\epsilon + \beta(1-\epsilon)$$

设:$p = P(Y=0)$,则:$H(Y) = -p\log p - (1-p)\log(1-p) \quad \cdots(1)$
而计算可知:

$$H(Y|X=0) = H(Y|X=1) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$$

因此:

$$H(Y|X) = P(X=0)H(Y|X=0) + P(X=1)H(Y|X=1) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$$

设:

$$f(x) = -x\log x - (1-x)\log(1-x)$$

则:

$$H(Y) = f(p) \quad H(Y|X) = f(\epsilon) \quad (1-\beta)(1-\epsilon) + \beta\epsilon$$

因此：

$$I(X;Y) = f((1-\beta)(1-\epsilon) + \beta\epsilon) - f(\epsilon)$$

2. A function $f(x)$ is called convex over (a,b) if for any $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2).$$

(a) Prove that $-\log x$ is a convex function.
(b) Prove Jensen's inequality:

$$\mathbb{E}[f(x)] \ge f(\mathbb{E}[x]),$$

where $f(x)$ is a convex function, and assume the probability distribution is discrete.
(c) Using Jensen's inequality, prove two properties about the relative entropy $D(P\|Q)$.

Solution

(a) $\forall a, b \in (0,\infty), a < b$,即要证:$\forall\lambda \in (0,1)$,$F(\lambda) := f(\lambda a + (1-\lambda)b) - \lambda f(a) - (1-\lambda)f(b) \le 0$,其中$f(x) = -\log x$.注意到:$F(0) = F(1) = 0$;而$F'(\lambda) = (a-b)f'(\lambda a + (1-\lambda)b) - f(a) + f(b) = -\log\frac{b}{a} - \frac{a-b}{(a-b)\lambda + b}$,上式关于λ单调递增,而$F'(0) =$

$-\log\frac{b}{a}-\frac{a-b}{b}<0$,$F'(0)=-\log\frac{b}{a}-\frac{a-b}{a}>0$(由不等式$x\geq\log x+1$,代入$x=\frac{a}{b},\frac{b}{a}$即可得证,而该不等式求导即可证明),容易说明F'(x)在$[a,b]$上是连续的,故F在$[a,b]$上先递减后递增,由端点处的取值进而得知$F(\lambda)\leq 0$,从而证明了结论

(b) 对k个取值的离散分布$P(x)$,有:$\mathbb{E}[f(x)]=\sum\limits_{i=1}^{k}p_if(x_i),\mathbb{E}[x]=\sum\limits_{i=1}^{k}p_ix_i$,即要证:$f(\sum\limits_{i=1}^{k}p_ix_i)\leq\sum\limits_{i=1}^{k}p_if(x_i)$,k=1,2时显然成立,假设k=m时成立,则k=m+1时,由于$1-p_{k+1}=\sum\limits_{i=1}^{k}p_i$,$f(\sum\limits_{i=1}^{k+1}p_ix_i)=f((1-p_{k+1})\sum\limits_{i=1}^{k}\frac{p_i}{1-p_{k+1}}x_i+p_{k+1}x_{k+1})\leq(1-p_{k+1})f(\sum\limits_{i=1}^{k}\frac{p_i}{1-p_{k+1}}x_i)+p_{k+1}f(x_{k+1})\leq(1-p_{k+1})\sum\limits_{i=1}^{k}\frac{p_i}{1-p_{k+1}}f(x_i)+p_{k+1}f(x_{k+1})\leq\sum\limits_{i=1}^{k+1}p_if(x_i)$,得证

(c)
即要证$D(P||Q)\geq 0$，等号成立当且仅当$P=Q$.设$f(t)=-\log t$，随机变量X的密度函数是P,考虑随机变量Q(X)/P(X),由Jensen不等式：

$$D(P||Q)=\mathbb{E}\left[-\log\frac{Q(X)}{P(X)}\right]\geq-\log\mathbb{E}\left[\frac{Q(X)}{P(X)}\right]=-\log 1=0$$

取等当且仅当Jensen取等.由-log的严格凸性,即要求P/Q=const,又因为P,Q均为分布函数,所以P=Q

3. Consider linear models where the target variable y follows a Gaussian distribution:

$$p(y\mid\mathbf{x},\mathbf{w},\sigma^2)=\mathcal{N}(y\mid\mathbf{w}^\top\mathbf{x},\sigma^2).$$

The mean is a linear combination of the features:

$$\mu=\mathbf{w}^\top\mathbf{x}=w_0+w_1x_1+\cdots+w_Dx_D.$$

Assuming the parameter $\mathbf{w}$ satisfies the following Gaussian prior distribution:

$$p(\mathbf{w})=\mathcal{N}(\mathbf{w}\mid\mathbf{0},\lambda^{-1}\mathbf{I}),$$

(a) Find the posterior distribution for $\mathbf{w}$, given that the dataset $\mathcal{D}$ consists of N points $\{(y_1,\mathbf{x}_1),\ldots,(y_N,\mathbf{x}_N)\}$.
(b) Find the loss function from MAP estimation of the posterior probability.

Solution
(a)

$$\begin{aligned}p(\mathcal{D}|\mathbf{w},\sigma^2)&=\prod_{n=1}^{N}\mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n,\sigma^2)\\&=\prod_{n=1}^{N}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(y_n-\mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2}\right)\\&=(2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2\right)\\p(\mathbf{w})&=\left(\frac{\lambda}{2\pi}\right)^{D/2}\exp\left(-\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right)\end{aligned}$$

因此,后验分布:

$$
\begin{aligned}
p(\mathbf{w} \mid \mathcal{D}) &\propto p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w}) \\
&= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2\right) \\
&\quad \cdot \left(\frac{\lambda}{2\pi}\right)^{D/2} \exp\left(-\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w}\right] - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right)
\end{aligned}
$$

其中,$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$,$\mathbf{y} = (y_1, \ldots, y_N)^T$
(8)中指数部分的二次项为:$-\frac{1}{2}\mathbf{w}^T\left[\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right]\mathbf{w}$,一次项为:$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{X}\mathbf{w}$
由此可知,$p(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu, \mathbf{\Sigma})$
其中,$\mathbf{\Sigma} = \left[\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right]^{-1}$, $\mu = \frac{1}{\sigma^2}\mathbf{\Sigma}\mathbf{X}^T\mathbf{y}$

(b)

$$
\begin{aligned}
\hat{\mathbf{w}}_{\mathrm{MAP}} &= \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathcal{D}) \\
&= \arg\max_{\mathbf{w}} \left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right] \\
&= \arg\min_{\mathbf{w}} \left[\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + \lambda\sigma^2\,\mathbf{w}^T\mathbf{w}\right]
\end{aligned}
$$

故损失函数为:

$$
\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + \lambda\sigma^2\,\mathbf{w}^T\mathbf{w}
$$

4. Consider a discrete probability distribution with $|A|$ outcomes. Find the probability distribution that maximizes the entropy.

Solution
$H(P) = -\sum_{i=1}^{|A|} p_i \log p_i$,其中$\sum_{i=1}^{|A|} p_i = 1$.由Lagrange乘子法,$L = -\sum_i p_i \log p_i + \lambda\left(\sum_i p_i - 1\right)$,则$\frac{\partial \mathcal{L}}{\partial p_i} = -\log p_i - 1 + \lambda = 0$,
解得$p_i = e^{\lambda-1}$.又因为$\sum_i p_i = |A|e^{\lambda-1} = 1$,所以$e^{\lambda-1} = \frac{1}{|A|}$,进而$p_i = \frac{1}{|A|}, i = 1, \ldots, |A|$.所以最大熵分布是离散的均匀分布