ML homework1

October 3, 2024

1 Q1. Maximum likelihood Estimates

1.1 Q1.1

we need to find the expectation of the function.

$$
\begin{aligned}
E[x] &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x \cdot e^{-\lambda}}{x!} \\
&= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{x \cdot \lambda^{x-1}}{x!} \\
&= \lambda e^{-\lambda} \cdot (\sum_{x=0}^{\infty} \frac{\lambda^x}{x!})' \\
&= \lambda e^{-\lambda} \cdot e^{\lambda} \\
&= \lambda
\end{aligned}
$$

1.2 Q1.2

Since the data $X_1, X_2, \ldots X_n$ are independent

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Useing maximum likelihood estimate, construct estimate function.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Taking the log of function to simplify calculations

$$ln(L(\lambda)) = (\sum_{i=1}^{n} x_i) ln\lambda - n\lambda - \sum_{i=1}^{n} ln(x!)$$

to find the minimize point , take the partial derivative of the function

$$\frac{\partial ln(L(\lambda))}{\partial\lambda} = \frac{\sum x_i}{\lambda} - n = 0$$

then we can get the minimize λ

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

2 Q2. Deriving Naïve Bayes

2.1 Q2.1

for dataset $\{x_i, y_i\}_{i=1}^n$ we know that the prior is the Bernoulli distribution. using MLE to find the formal p

$$L(p) = \prod_{i=1}^{n} P_Y(y_i) = \prod p^{y_i} \cdot (1-p)^(1-y_i)$$

$$ln(L(p)) = \sum_{i=1}^{n}[y_i lnp + (1-y_i)ln(1-p)]$$

$$\frac{\partial ln(L(p))}{\partial p} = \sum(\frac{y_i}{p} + \frac{1-y_i}{p-1}) = 0$$

$$\frac{\sum y_i}{p} = \frac{n}{1-p} - \frac{\sum y_i}{1-p}$$

we can get the formal p from above equation

$$\sum y_i = np \qquad\qquad p = \frac{\sum_{i=1}^{n} y_i}{n}$$

2.2 Q2.2

we can obtain two subset based on different y label
subset 0:$\{x_i|y_i = 0\}$ subset 1: $\{x_i|y_i = 1\}$
Assume each subset have k elements.The likelihood function is:

$$L(\mu_y, \sigma_y^2) = \prod_{i:y=y_i}^{k} \frac{1}{\sqrt{2\pi\sigma_y^2}} \cdot e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}}$$

$$ln(L(\mu_y, \sigma_y^2)) = -\frac{k}{2}ln(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2}\sum(x_i^2 - 2x_i\sigma_y + \sigma_y^2) = 0$$

Take the partial derivative of the function using σ_y^2 and μ_y respectively then we obtain two equations:

$$k\sigma_y^2 + \sum_{i=1}^{k} x_i^2 - 2(\sum_{i=1}^{k} \sigma_y) = 0$$

$$2k\mu_y - 2(\sum_{i=1}^{k} x_i) = 0$$

find the arguments of each subset:

$$\mu = \frac{1}{k} \cdot \sum_{i=1}^{k} x_i \qquad \sigma = \frac{1}{k}(x_i - y_i)^2$$

3 Q3. Ridge Regression

3.1 Q3.1

$$\min \frac{1}{n}||X \cdot \omega - y||^2 + \lambda||\omega||^2$$

3.2 Q3.2

$$\begin{aligned} J(\omega) &= \frac{1}{n}||X \cdot \omega - y||^2 + \lambda||\omega||^2 \\ &= \frac{1}{n}(X \cdot \omega_d - y)^T(X \cdot \omega_d - y) + \lambda\omega^T\omega \\ &= \frac{1}{n}(\omega^T X^T X\omega - 2y^T X\omega + y^T y) + \lambda\omega^T\omega \end{aligned}$$

Take the partial derivative of the function using ω_d

$$\frac{\partial J(\omega)}{\partial\omega} = \frac{1}{n}(2TX^T X\omega - 2X^T y) + 2\lambda\omega = 0$$

$$X^T y - X^T X\omega = n\lambda\omega$$

$$(X^T X + n\lambda I)\omega = X^T y$$

3.3 Q3.3

racall that:

$$\min \frac{1}{n}||X \cdot \omega - y||^2 + \lambda||\omega||^2$$

$$(X^T X + n\lambda I)\omega = X^T y$$

$\rightarrow$

$$\omega = (X^T X + n\lambda I)^{-1} X^T y$$

we will use above two equations to find the solution

$$\begin{aligned}
J(\omega) &= \frac{1}{n}(\omega^T X^T X\omega - 2y^T X\omega + y^T y) + \lambda\omega^T\omega \\
&= \frac{1}{n}(\omega^T X^T X\omega - 2y^T X\omega + y^T y + n\lambda\omega^T\omega) \\
&= \frac{1}{n}[\omega^T(X^T X\omega + n\lambda\omega) - 2y^T X\omega + y^T y] \\
&= \frac{1}{n}[\omega^T X^T y - 2y^T X(X^{-1}y + \frac{1}{n\lambda}X^T y) + y^T y] \\
&= \frac{1}{n}[\omega^T X^T y - y^T y - \frac{2}{n\lambda}y^T X^T y]
\end{aligned}$$

3.4 Q3.4

recall that :

$$\omega = (X^T X + n\lambda I)^{-1} X^T y$$

when $(X^T X + n\lambda I)$ has full rank, means $(X^T X + n\lambda I)^{-1}$ exist. The critical point is unique.
when $\lambda \succ 0$, $n\lambda I$ ensures $(X^T X + n\lambda I)$ is **positive definite**, then this matrix is invertible, that's say the critical point is unique.
Summary: $\lambda \succ 0$, and matrix X is real matrix.

4 Q4. Optimal predictor

4.1 Q4.1

assume that:

$$M(x, y) = E_{(X,Y)\ P_{XY}}[(f(x) - y)^2]$$

take the partial derivative

$$\frac{\partial M(x, y)}{\partial f(x)} = E_{y|X=x}[2f(x) - 2y] = 0$$

the variable of the above function is y. that's means that:

$$2f(x) - 2E_{y|X=x}[Y] = 0$$

$$f(x) = E_{y|X=x}[Y] = 0$$

4.2 Q4.2

$$E[|f(X) - Y|] = E[E[|f(x) - Y|X = x]]$$

for each given x ,we need to minimize $E[|f(x) - Y|X = x]$

$$\begin{aligned} L(x) &= E[|f(x) - y||X = x] \\ &= \int_{-\infty}^{+\infty} |f(x) - y| \cdot p(y|x)\mathrm{d}y \end{aligned}$$

we get that:

$$\begin{aligned} \frac{\mathrm{d}L(x)}{\mathrm{d}f(x)} &= \int_{-\infty}^{f(x)} p(y|x)\mathrm{d}y - \int_{f(x)}^{+\infty} p(y|x)\mathrm{d}y \\ &= 2P(Y \leq f(x)|X = x) - 1 \end{aligned}$$

That's means:

$$P(Y \leq f(x)|X = x) = 0.5$$

So $f^*(x)$ is the median number of Y