中 国 作 家 协 会

STA414. Assignment 1.

#1.

• Solution: $cov(x,y) = E[(x-E(x))(y-E(y))]$
$= E[xy - E(x)y - xE(y) + E(x)E(y)]$
$= E(xy) - E(x)E(y)$

(2) Since $x \perp y$, $p(xy) = p(x)p(y)$
so $E(xy) = \sum\sum p(xy)\,xy$
$= \sum\sum p(x)p(y)\,xy$
$= E(x)E(y)$

Therefore $cov(x,y) = 0$.

• Proof: $E(x) = \mu$, $Var(x) = \sigma^2$. Want to show $E(x_n x_m) = \mu^2 + I_{nm}\sigma^2$

~~Since the expectation~~ $n=m$

① Suppose $x_n = x_m$, then $E(x_n^2) = \mu^2 + \sigma^2$
We know that the second moment of a Gaussian distribution is

$$E(x^2) = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\,x^2\,dx = \mu^2 + \sigma^2 \qquad (*)$$

② Suppose ~~$x_n \neq x_m$~~ $n \neq m$, $x_n \neq x_m$, so $E(x_n x_m) = E(x_n)E(x_m)$ b/c independence

Then $E(x_n x_m) = \mu \cdot \mu = \mu^2$ (**)

Consider (*) & (**) together, we now have

$$E(x_n x_m) = \mu^2 + I_{nm}\sigma^2 \quad \text{with} \quad I_{nm} = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{otherwise.} \end{cases}$$

• Proof: joint distribution $p(x,y)$
To prove equation one the key is to expand expectation to the form of integrals.

$$E_y E_x[x|y] = E_y\left(\int x\,p(x|y)\,dx\right) = \int\left(\int x\,p(x|y)\,dx\right)p(y)\,dy$$
$$= \iint x\,p(x,y)\,dx\,dy \quad \text{by product rule: } p(x|y)p(y) = p(x,y)$$
$$= E(x)$$

中国作家协会

To prove the second equation,

we write:

$$Var(x) = E(x^2) - [E(x)]^2$$

and we apply the result of first equation:

$$E_y[var_x[x|y]] = E_y[E_x[x^2|y] - E_x[x|y]^2] \quad ①$$

$$var_y[E_x[x|y]] = E_y[E_x[x|y]^2] - E_y[E_x[x|y]]^2 \quad ②$$

$$① + ② = E[x^2] - E_y[E_x[x|y]]^2 + E_y[E_x[x|y]^2] - (E_x[x])^2$$

$$= Var(x)$$

#2. $p(\text{mistake}) = \int_{R_1} p(x, C_2)\,dx + \int_{R_2} p(x, C_1)\,dx$

Proof: Want to show $p(\text{mistake}) \leq \int [p(x, C_1)\, p(x, C_2)]^{\frac{1}{2}}\,dx$

First we might want to use the product rule of probability to rewrite this:

$$p(\text{mistake}) = \int_{R_1} p(C_2|x)p(x)\,dx + \int_{R_2} p(C_1|x)p(x)\,dx$$

Since we want to minimize the probability of mistake, i.e. the value of x should make $p(C_2|x) \leq p(C_1|x)$ in R_1 and vice versa in R_2.

Then $p(\text{mistake}) = \int_{R_1} p(C_2|x)p(x)\,dx + \int_{R_2} p(C_1|x)p(x)\,dx$

$$\leq \int_{R_1} [p(C_2|x)\,p(C_1|x)]^{\frac{1}{2}}\,p(x)\,dx + \int_{R_2} [p(C_1|x)\,p(C_2|x)]^{\frac{1}{2}}\,p(x)\,dx$$

$$= \int_{R_1+R_2} [p(C_2|x)\,p(x)\,p(C_1|x)\,p(x)]^{\frac{1}{2}}\,dx \quad \text{\# same integrand}$$

$$= \int [p(x, C_2)\,p(x, C_1)]^{\frac{1}{2}}\,dx \quad \text{\# use product rule again}$$

中 国 作 家 协 会

#3.

Proof: pdf: $f = \binom{N}{m} \mu^m (1-\mu)^{N-m}$

$$E(m) = \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$= \sum_{m=1}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$= \sum_{m=1}^{N} N \binom{N-1}{m-1} \mu^m (1-\mu)^{N-m} \qquad \text{b/c } a\binom{b}{a} = b\binom{b-1}{a-1}$$

$$= N\mu \sum_{m=1}^{N} \binom{N-1}{m-1} \mu^{m-1} (1-\mu)^{(N-1)-(m-1)}$$

$$= N\mu (\mu + (1-\mu))^{N-1} \qquad \text{by binomial theorem.}$$

$$= N\mu$$

$$Var(m) = E(m^2) - [E(m)]^2$$

$$E(m^2) = \sum_{m=0}^{N} m^2 \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$= \sum_{m=0}^{N} mN \binom{N-1}{m-1} \mu^m (1-\mu)^{N-m}$$

$$= N\mu \sum_{m=0}^{N} m \binom{N-1}{m-1} \mu^{m-1} (1-\mu)^{N-m} \qquad \text{let } \begin{matrix} N-1 = s \\ m-1 = t \end{matrix}$$

$$= N\mu \sum_{t=0}^{s} (t+1) \binom{s}{t} \mu^t (1-\mu)^{s-t}$$

$$= N\mu \left(\sum_{t=0}^{s} t \binom{s}{t} \mu^t (1-\mu)^{s-t} + \sum_{t=0}^{s} \binom{s}{t} \mu^t (1-\mu)^{s-t} \right)$$

$$= N\mu \left((N-1) \sum_{t=1}^{s} \binom{s-1}{t-1} \mu^{t-1} (1-\mu)^{(s-1)-(t-1)} + \sum_{t=0}^{s} \binom{s}{t} \mu^t (1-\mu)^{s-t} \right)$$

$$= N\mu \left((N-1)\mu (\mu + 1 - \mu)^{s-1} + (\mu + 1 - \mu)^s \right)$$

$$= N\mu (N\mu - \mu + 1)$$

$$= N^2\mu^2 - N\mu^2 + N\mu$$

So $Var(m) = E(m^2) - [E(m)]^2 = N^2\mu^2 - N\mu^2 + N\mu - N^2\mu^2 = N\mu(1-\mu)$

中 国 作 家 协 会

#4. Proof: Sps Σ is $n \times n$ matrix (i.e. $D=n$)

Σ is symmetric, one can write it as a decomposition:

$\Sigma = U \Lambda U^T$ where U orthogonal, Λ diagonal with $\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \lambda_n \end{pmatrix}$

$$\Sigma x = U \Lambda U^T x = U \Lambda \begin{pmatrix} u_1^T x \\ \vdots \\ u_n^T x \end{pmatrix} = U \begin{pmatrix} \lambda_1 u_1^T x \\ \vdots \\ \lambda_n u_n^T x \end{pmatrix}$$

essentially we want to prove this

$$= \sum_{i=1}^{n} \lambda_i u_i^T x \cdot u_i$$

$$= \left(\sum_{i=1}^{n} \lambda_i u_i u_i^T \right) x$$

Hence $\Sigma = \sum_{i=1}^{D} \lambda_i u_i u_i^T$

Invert $\Sigma = U \Lambda U^T$ from both sides. note that U is orthogonal

i.e. $U^{-1} = U^T$

so $\Sigma^{-1} = U \Lambda^{-1} U^T$

$$= \sum_{i=1}^{n} \frac{1}{\lambda_i} u_i u_i^T$$

b/c Λ^{-1} is a diagonal matrix whose diagonal elements are inverses of their counterparts in diagonal matrix Λ.

7

∎

中 国 作 家 协 会

#5.

Proof: Note that the coefficient before exponential part $\exp(-\frac{|x|^q}{2\sigma^2})$ must "normalize" the distribution.

So we only calculate $\int_{-\infty}^{\infty} \exp(-\frac{|x|^q}{2\sigma^2})\,dx$

$= 2\int_0^{\infty} \exp(-\frac{|x|^q}{2\sigma^2})\,dx$ let $u=\frac{x^q}{2\sigma^2}$, so $du=\frac{1}{2\sigma^2} q x^{q-1}$

$= 2\int_0^{\infty} \frac{2\sigma^2(2\sigma^2 u)^{\frac{1-q}{q}}}{q} \exp(-u)\,du$

$= \frac{2(2\sigma^2)^{\frac{1}{q}}}{q}\int_0^{\infty} u^{\frac{1-q}{q}}\exp(-u)\,du = \frac{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})}{q}$ because $\Gamma(x)=\int_0^{\infty} u^{x-1}e^{-u}\,du$

Hence this times the term $\frac{q}{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})}$ is equal to 1.

Normalization proved. Also $p>0$ because exponential is nonnegative. with integral part = 1. So it's a valid distribution.

- When $q=2$, $p(x|\sigma^2,2)=\frac{1}{(2\sigma^2)^{\frac{1}{2}}\Gamma(\frac{1}{2})}\exp(-\frac{(x-0)^2}{2\sigma^2})$

$= \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-0)^2}{2\sigma^2})$ since $\Gamma(\frac{1}{2})=\sqrt{\pi}$

It's Gaussian.

(this can be proved by applying a transformation to change variables to polar expressions).

- target variable $t=y(x,w)+\varepsilon$, ε is random noise

Log likelihood function over w, σ^2?

observed data $X=\{x_1,\cdots,x_N\}$

$t=(t_1,\cdots,t_N)^T$

(see next page)

中 国 作 家 协 会

~~$t = y(x, w)$~~ $\xi = t - y(x, w)$

Given noise variable, the conditional distribution of target variable t and input x, w is

$$p(t|x, w, \sigma^2) = \frac{q}{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})} \exp\left(-\frac{|t - y(x, w)|^q}{2\sigma^2}\right)$$

For the derivation of Log likelihood function, we take logarithm on both sides of the product of such conditional probability. The process is quite tedious with long terms of constants:

$$L = \text{重} \ln \prod_{i \in \{1, \cdots, N\}} \left\{ \frac{q}{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})} \exp\left(-\frac{|t_i - y(x_i, w)|^q}{2\sigma^2}\right) \right\}$$

$$= \ln \left\{ \left\{ \frac{q}{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})} \right\}^N \exp\left(\sum_{i=1}^{N} \frac{-|t_i - y(x_i, w)|^q}{2\sigma^2} \right) \right\}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} |t_i - y(x_i, w)|^q + N \ln \left\{ \frac{q}{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})} \right\}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} |t_i - y(x_i, w)|^q \text{事} - \frac{N}{q} \ln 2\sigma^2 + N\left(\ln q - \frac{1}{q}\ln 2 - \ln\Gamma(\frac{1}{q})\right)$$

with fixed q, this term is a constant real number, has nothing to do with any of x, w or σ^2.

6

中 国 作 家 协 会

6. expected L_q loss function

$$E[L_q] = \iint |y(x) - t|^q p(x,t)\, dt\, dx$$

(w). Proof: Let $y(x) \perp x$ so that $p(x,t) = p(t|x)$

so $$\frac{\partial}{\partial y(x)} E[L_q] = \frac{\partial}{\partial y(x)} \iint |y(x)-t|^q p(t|x)\, dt\, dx$$

$$= \int q\,|y(x)-t|^{q-1} p(t|x)\, dt$$

$$= 0$$ (since minimized, and to simplify. I didn't check the 2nd derivative)

$$\Rightarrow \int_{-\infty}^{\infty} |y(x)-t|^{q-1} p(t|x)\, dt = 0 \qquad (*)$$

Since we want to check the median (and its relation with probability mass), so we split (*) into two parts: ~~set the same time $q=1$~~

~~$\int_{y(x)}^{\infty} p(t|x)\, dt = \int_{-\infty}^{y(x)} p(t|x)\, dt$~~

$$\int_{y(x)}^{\infty} |y(x)-t|^{q-1} p(t|x)\, dt = \int_{-\infty}^{y(x)} |y(x)-t|^{q-1} p(t|x)\, dt$$

where left part for $t \geq y(x)$, right part for $t < y(x)$.

Plug in $q=1$, $\int_{y(x)}^{\infty} p(t|x)\, dt = \int_{-\infty}^{y(x)} p(t|x)\, dt$

so $y(x)$ is the conditional median for $\forall x$.

■

中 国 作 家 协 会

b). Proof: as $q \to 0$, $|y(x)-t|^q \to 1$ ~~exp~~ except when $y(x)-t=0$, since 0^0 doesn't exist.

Therefore $\int_{-\infty}^{\infty} |y(x)-t|^q p(t|x)\,dt \to \int_{-\infty}^{\infty} p(t|x)\,dt \to 1$

except some points near $y(x)-t=0$.
At these points. the integral above is less than 1
because these points should be excluded from the
integra~~tion~~ation.
Thus. in order to minimize the value of this integral,
we need to have $y(x)=t$ as frequent as possible.
That is to say, to have this with the largest value
of $p(t|x)$, i.e. the most possible scenario.
Therefore, it's given by the conditional mode.

(# problem 7 see the back of this page)

7.

Proof: $\forall v \in \mathbb{R}^m$

$$(\Phi(\Phi^T\Phi)^{-1}\Phi^T)^2 v$$

$$= \Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^T v$$

$$= \Phi(\Phi^T\Phi)^{-1}(\Phi^T\Phi)(\Phi^T\Phi)^{-1}\Phi^T v$$

$$= \Phi(\Phi^T\Phi)^{-1}\Phi^T v \in \text{Image}(\Phi).$$

So the matrix takes any v and projects it onto the space spanned by the columns of Φ.

Since $w^* = (\Phi^T\Phi)^{-1}\Phi^T t$

$$\Phi w^* = \Phi(\Phi^T\Phi)^{-1}\Phi^T t$$

Check for orthogonality:

Suppose ϕ_i is ith column of Φ.

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T\phi_i = \phi_i = I\phi_i$$

$$(\Phi w^* - t)^T\phi_i = (\Phi(\Phi^T\Phi)^{-1}\Phi^T - I)^T\phi_i$$

$$= t^T(I-I)\phi_i$$

$$= 0$$

So such matrix is is an orthogonal projection.

#8.

Proof: Know that $p(x|\eta) = h(x)g(\eta)\exp(\eta^T u(x))$

$$-\nabla \ln g(\eta) = E(u(x))$$

$$= g(\eta)\int h(x)\exp(\eta^T u(x))u(x)dx$$

$$= g(\eta)\int p(x|\eta)u(x)dx$$

10

So similarly,

$$-\nabla\nabla \ln g(\eta) = \int\{h(x)g(\eta)\exp(\eta^T u(x))u(x)u(x)^T dx + h(x)\nabla g(\eta)\exp(\eta^T u(x))u(x)\}dx$$

↓ split into two parts

$$= \int h(x)g(\eta)\exp(\eta^T u(x))u(x)u(x)^T dx + \int h(x)\nabla g(\eta)\exp(\eta^T u(x))u(x)dx$$

$$= g(\eta)\int h(x)\exp(\eta^T u(x))u(x)u(x)^T dx + \nabla g(\eta)\int h(x)\exp(\eta^T u(x))u(x)dx$$

$$= E(u(x)u(x)^T) - E(u(x))E(u(x)^T)$$ by equations above ↑

$$= cov(u(x), u(x)^T)$$

$$= cov(u(x))$$

↳ it's a vector here.

∎