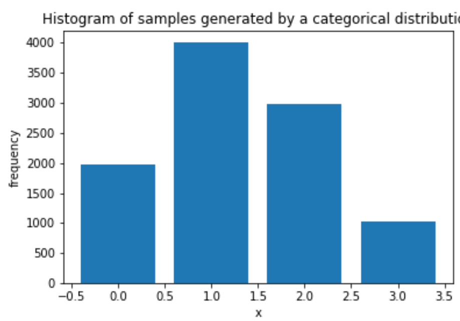
COMP 540 HW 01

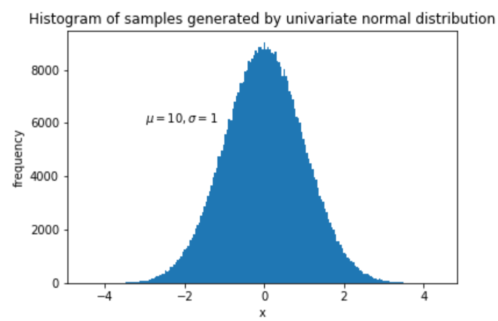
Lyu Pan (lp28), Yuhui Tong (yt30)

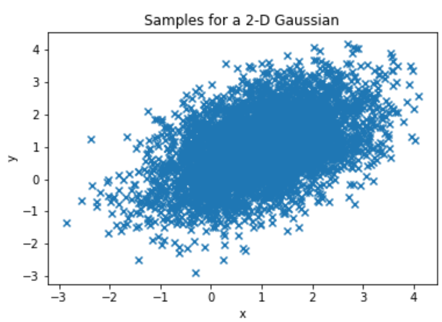
January 15, 2018

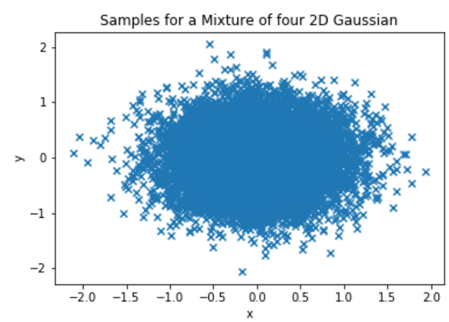
0. Background Refresher

0.0 Samplers









0.1 Prove two independent Poisson random variables are also Poisson variable.

Proof:

Given two independent random variables $X_1 \sim P(\lambda_1)$ and $X \sim P(\lambda_2)$. i.e.,

$$P(X_1 = m) = e^{-\lambda_1}\frac{\lambda_1^m}{m!},\ m = 1, 2, ... \tag{1}$$

and

$$P(X_2 = n) = e^{-\lambda_2}\frac{\lambda_2^n}{n!},\ n = 0, 1, 2, ... \tag{2}$$

Denote sum of them are

$$X = X_1 + X_2 \tag{3}$$

then the probability distribution of X is:

$$\begin{aligned}
P(X=k) &= \sum_{m+n=k} P(X_1=m, X_2=n) \\
&\overset{X_1,X_2 indep.R.V.}{=} \sum_{m+n=k} P(X_1=m)P(X_2=n) \\
&= \sum_{m+n=k} e^{-\lambda_1}\frac{\lambda_1^m}{m!}e^{-\lambda_2}\frac{\lambda_2^n}{n!} \\
&= \frac{1}{k!}e^{-(\lambda_1+\lambda_2)}\sum_{m+n=k}\frac{k!}{m!n!}\lambda_1^m\lambda_2^n \\
&= \frac{1}{k!}e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k \qquad (4)
\end{aligned}$$

that is, $X = X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$. Q.E.D.

0.2 Proof question 2

$$\begin{aligned}
p(x_1,x_0) &= p(x_1|x_0)p(x_0) \\
&= \alpha e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}\alpha_0 e^{-\frac{(x_0-\mu_0)^2}{2\sigma_0^2}} \qquad (5)
\end{aligned}$$

The probability distribution of X_1 is,

$$\begin{aligned}
p(x_1) &= \int_{-\infty}^{\infty} p(x_1,x_0)dx_0 \\
&= \int_{-\infty}^{\infty} \alpha e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}\alpha_0 e^{-\frac{(x_0-\mu_0)^2}{2\sigma_0^2}}dx_0 \\
&= D\int_{-\infty}^{\infty} e^{-A[x_0-B]^2+C}dx_0 \qquad (6)
\end{aligned}$$

where A, B, C, D

$$D = \alpha\alpha_0 e^{-\frac{\mu_0^2}{2\sigma_0^2}}e^{-\frac{x_1^2}{2\sigma^2}} \qquad (7)$$

$$A = \frac{\sigma_0^2+\sigma^2}{2\sigma_0^2\sigma^2}, \qquad (8)$$

$$B = \frac{x_1\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2}, \qquad (9)$$

$$C = \frac{(x_1\sigma_0^2+\mu_0\sigma^2)^2}{2\sigma_0^2\sigma^2(\sigma_0^2+\sigma^2)}. \qquad (10)$$

Making the substitution $t = \sqrt{2A}(x_0 - B)$ gives

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-A[x_0-B]^2+C} dx_0 &= \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}+C} d\frac{t}{\sqrt{2A}} \\ &= \frac{1}{\sqrt{2A}} e^{-C} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \\ &= \sqrt{\frac{\pi}{A}} e^{-C} \end{aligned} \tag{11}$$

where in the second step we used $\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$.

Substitute the above equation to the Eq. ():

$$\begin{aligned} p(x_1) &= D\sqrt{\frac{\pi}{A}} e^{-C} \\ &= \alpha\alpha_0 e^{-\frac{\mu_0^2}{2\sigma_0^2}} e^{-\frac{x_1^2}{2\sigma^2}} \sqrt{\frac{\pi}{A}} e^{-C} \end{aligned} \tag{12}$$

0.3 question 4 eigenvalues

$$\boldsymbol{A} = \begin{bmatrix} 13 & 5 \\ 2 & 4 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{AX} &= \lambda \boldsymbol{X} \\ (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{X} &= 0 \\ \begin{bmatrix} 13-\lambda & 5 \\ 2 & 4-\lambda \end{bmatrix} \boldsymbol{X} &= 0 \end{aligned}$$

$\lambda_1 = 3,\ \boldsymbol{X}_1 = \begin{pmatrix} -1/2 \\ 1 \end{pmatrix}$

$\lambda_2 = 14,\ \boldsymbol{X}_2 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$

0.4 question 5, matrix multiplication

$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$, we have $(A+B)^2 \neq A^2 + 2AB + B^2$

$A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$, $A, B \neq 0$ and we have $AB = 0$

0.5 question 6

$$\begin{aligned} A^T A &= (I - 2uu^T)^T (I - 2uu^T) \\ &= (I - 2uu^T)(I - 2uu^T) \\ &= I - 2uu^T - 2uu^T + 4u(u^T u)u^T \\ &= I \end{aligned} \tag{13}$$

0.6 convex function

0.7 entropy of categorical distribution

The entropy of a categorical distribution on K values is

$$H(p) \quad = \quad -\sum_{i=1}^{K} p_i \log(p_i), \tag{14}$$

with constraint that

$$\sum_{i=1}^{K} p_i = 1. \tag{15}$$

Using Lagrange Multiplier, one can combine the above two equations into:

$$L(p, \lambda) = -\sum_{i=1}^{K} p_i \log(p_i) + \lambda(\sum_{i=1}^{K} p_i - 1). \tag{16}$$

Taking derivative of all unknown variables gives:

$$\frac{\partial L}{\partial p_i} = -(\log p_i + 1) + \lambda = 0, \;\; i = 1, 2, ... \tag{17}$$

Substituting p_i with λ yields

$$p_i = \frac{1}{K}, \;\; i = 1, 2, ..., K. \tag{18}$$

Q.E.D.

1. Locally weighted linear regression.

1.1 Expression of $J(\theta)$

Define $\boldsymbol{X}$, $\boldsymbol{W}$ in the following way:

$$\boldsymbol{X} = \begin{bmatrix} ----\left(x^{(1)}\right)^T ---- \\ ----\left(x^{(2)}\right)^T ---- \\ \vdots \\ ----\left(x^{(m)}\right)^T ---- \end{bmatrix} \Longleftrightarrow \boldsymbol{X}_{i,j} = (x^{(i)})_j \tag{19}$$

$$\boldsymbol{W} = \begin{bmatrix} w^{(1)} & & & \\ & w^{(2)} & & \\ & & \ddots & \\ & & & w^{(m)} \end{bmatrix} \Longleftrightarrow \boldsymbol{W}_{i,j} = \delta_{i,j} w^{(i)} \tag{20}$$

Substituting $\boldsymbol{X}$, $\boldsymbol{W}$ into $(\boldsymbol{X}\theta - \boldsymbol{y})^T \boldsymbol{W} (\boldsymbol{X}\theta - \boldsymbol{y})$ yields

$$
\begin{aligned}
(\boldsymbol{X}\theta-\boldsymbol{y})^T\boldsymbol{W}(\boldsymbol{X}\theta-\boldsymbol{y}) &= \sum_{i,j}\left[(\boldsymbol{X}\theta-\boldsymbol{y})^T\right]_{1,i}\boldsymbol{W}_{i,j}(\boldsymbol{X}\theta-\boldsymbol{y})_{j,1} \\
&= \sum_{i,j}\delta_{i,j}w^{(i)}\left[(\boldsymbol{X}\theta-\boldsymbol{y})^T\right]_{1,i}(\boldsymbol{X}\theta-\boldsymbol{y})_{j,1} \\
&= \sum_{i}w^{(i)}\left[(\boldsymbol{X}\theta-\boldsymbol{y})_{i,1}\right]^2 \\
&= \sum_{i}w^{(i)}\left[\left(x^{(i)}\right)^T\theta-y^{(i)}\right]^2 \\
&= \sum_{i}w^{(i)}\left[\theta x^{(i)}-y^{(i)}\right]^2.
\end{aligned}
$$

So $J(\theta)=\frac{1}{2}(\boldsymbol{X}\theta-\boldsymbol{y})^T\boldsymbol{W}(\boldsymbol{X}\theta-\boldsymbol{y})$. Q.E.D.

1.2 Closed formed solution of θ

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial\theta} &= \frac{1}{2}\frac{\partial}{\partial\theta}\left[(\boldsymbol{X}\theta-\boldsymbol{y})^T\boldsymbol{W}(\boldsymbol{X}\theta-\boldsymbol{y})\right] \\
&= \frac{1}{2}\frac{\partial}{\partial\theta}tr\left[\boldsymbol{y}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y}^T\boldsymbol{W}\boldsymbol{y}-\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\theta}+\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\right] \\
&= \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\theta}
\end{aligned}
$$

where in the third line we used the following two equations[1]:

$$
\frac{\partial tr(\boldsymbol{A}^T\boldsymbol{B})}{\partial\boldsymbol{A}}=\boldsymbol{B}^T \tag{21}
$$

and

$$
\frac{\partial tr(\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{A})}{\partial\boldsymbol{A}}=\boldsymbol{B}\boldsymbol{A}+\boldsymbol{B}^T\boldsymbol{A} \tag{22}
$$

Therefore the closed form solution for θ is

$$
\theta=(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}. \tag{23}
$$

[1] Take the second equation for example:

$$
\begin{aligned}
\left[\frac{\partial tr(\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{A})}{\partial\boldsymbol{A}}\right]_{i,j} &= \frac{\partial A^T_{l,m}B_{m,n}A_{n,l}}{\partial A_{i,j}}=\frac{\partial A_{m,l}B_{m,n}A_{n,l}}{\partial A_{i,j}} \\
&= \delta_{m,i}\delta_{l,j}B_{m,n}A_{n,l}+A_{m,l}B_{m,n}\delta_{n,i}\delta_{l,j} \\
&= B_{i,n}A_{n,j}+A_{m,j}B_{m,i} \\
&= (\boldsymbol{B}\boldsymbol{A})_{i,j}+(\boldsymbol{B}^T\boldsymbol{A})_{i,j},
\end{aligned}
$$

hence leading to $\frac{\partial tr(\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{A})}{\partial\boldsymbol{A}}=\boldsymbol{B}\boldsymbol{A}+\boldsymbol{B}^T\boldsymbol{A}$

1.3 Batch gradient descent for locally weighted linear regression

The derivative of J_θ is:

$$\frac{\partial}{\partial\theta}J(\theta) = \sum_{i=1}^{m} w^{(i)}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}. \tag{24}$$

Therefore we have the following algorithm:

Algorithm 0: Batch gradient descent algorithm for locally weighted linear regression

Result: The estimated parameters θ for locally weighted linear regression

while θ *does not converge* **do**
  for *every* j **do**
    $\theta_j = \theta_j - \alpha \sum_{i=1}^{m} w^{(i)}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$;
  end
end

2. Properties of linear regression estimator.

2.1 Prove $E[\theta] = \theta^*$

Proof:

The following facts:

1. $y^{(i)} = \theta^{*T}x^{(i)} + \epsilon^{(i)}$,
2. $\epsilon^{(i)}, i = 1, 2, ..., m$ are *i.i.d.* of $N(0, \sigma^2)$,

indicate that:

given fixed arbitrary $x^{(i)}$ and fixed unknown parameter θ^*, $y^{(i)}$, $1 \leq i \leq m$ are *i.i.d. of* $N(\theta^{*T}x^{(i)}, \sigma^2)$

$$p(y^{(i)}|x^{(i)}, \theta^*) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y^{(i)}-\theta^{*T}x^{(i)})^2}{2\sigma^2}}. \tag{25}$$

The least-square estimate of θ^* is θ given by

$$\theta = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{26}$$

$$\overset{denote:(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\equiv\boldsymbol{A}}{=} \boldsymbol{A}\boldsymbol{y} \tag{27}$$

The expectation of θ is

$$E[\theta] = E[\boldsymbol{A}\boldsymbol{y}] \tag{28}$$

$$= E\begin{bmatrix} \sum_j A_{1,j}y^{(j)} \\ \sum_j A_{2,j}y^{(j)} \\ \vdots \\ \sum_j A_{d+1,j}y^{(j)} \end{bmatrix}. \tag{29}$$

Since expectations has the following property (regardless of the independence of Z_k):

$$E[Z_1 + Z_2 + ... + Z_l] = \sum_k Z_k, \tag{30}$$

we have

$$\begin{aligned}
E[\theta] &= E \begin{bmatrix} \sum_j A_{1,j} y^{(j)} \\ \sum_j A_{2,j} y^{(j)} \\ \vdots \\ \sum_j A_{d+1,j} y^{(j)} \end{bmatrix} = \begin{bmatrix} \sum_j A_{1,j} E[y^{(j)}] \\ \sum_j A_{2,j} E[y^{(j)}] \\ \vdots \\ \sum_j A_{d+1,j} E[y^{(j)}] \end{bmatrix} && (31) \\
&= \boldsymbol{A} \begin{bmatrix} E[y^{(1)}] \\ E[y^{(2)}] \\ \vdots \\ E[y^{(m)}] \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \theta^{*T} x^{(1)} \\ \theta^{*T} x^{(2)} \\ \vdots \\ \theta^{*T} x^{(m)} \end{bmatrix} && (32) \\
&= \boldsymbol{A}\boldsymbol{X}\theta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\theta^* && (33) \\
&= \theta^* && (34)
\end{aligned}$$

where in the second step the following equation is used: $E[y^{(i)}] = \theta^{*T} x^{(i)}$ (trivial to obtain as $y^{(i)}$ observe normal distribution). Therefore $E[\theta] = \theta^*$, implying that the estimation θ is unbiased. Q.E.D

2.2 Prove $Var(\theta) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2$

Denote $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, therefore $\theta = \boldsymbol{A}\boldsymbol{y}$.

$$\begin{aligned}
Var(\theta) &= Var(\boldsymbol{A}\boldsymbol{y}) && (35) \\
&= Var(\begin{bmatrix} \sum_j A_{1,j} y^{(j)} \\ \sum_j A_{2,j} y^{(j)} \\ \vdots \\ \sum_j A_{m,j} y^{(j)} \end{bmatrix}) = \begin{bmatrix} \sum_j A_{1,j} Var(y^{(j)}) \\ \sum_j A_{2,j} Var(y^{(j)}) \\ \vdots \\ \sum_j A_{m,j} Var(y^{(j)}) \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} Var(y^{(1)}) \\ Var(y^{(2)}) \\ \vdots \\ Var(y^{(m)}) \end{bmatrix}, && (36)
\end{aligned}$$

where in the second line we made used of the property of variance:

$$Var(\sum_k Z_k) = \sum Var(Z_k),\ Z_k\ is\ independent\ from\ each\ other. \tag{37}$$

Substituting $Var(y^{(i)}) = \sigma^2\ i = 1, 2, ..., m$, we have $Var(\theta) = \boldsymbol{A}\sigma^2 = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2$. Q.E.D.