Machine Learning

Siddharth Shah SBU ID-110957500

February 2017

1 Linear Algebra

1. Prove that $A \in R^{nxn}$ and A^T have the same eigenvalues

   Proof

   We know that for a non zero vector v, the eigen equations is represented as

   $$Av = \lambda v \tag{1}$$

   We can solve for the eigen values by solving the below equation

   $$Av - \lambda v = 0 \tag{2}$$

   $$(A - \lambda I)v = 0 \tag{3}$$

   where I is a nxn identity matrix. The above equation can have non zero v only when the determinant $|A - \lambda I| = 0$ and thus we need to solve for

   $$|A - \lambda I| = 0$$

   We also know that the $|A| = |A^T|$. Taking transpose on each side. We have the following equation

   $$|A - \lambda I|^T = 0$$

   which is the same as

   $$|A^T - \lambda I^T| = 0$$

   Since $I^T = I$

   $$|A^T - \lambda I| = 0$$

   Thus, we can say that the A and A^T have the same eigen values.

2. Let λ_i are the eigenvalues of $M \in R^{nxn}$. Determine the eigenvalues of $\alpha M + \beta I$, where I is the identity matrix, and $\alpha, \beta \in R$.

   Proof: Let $x \neq 0$ be an eigen vector, such that $Mx = \lambda_i x$

We can now consider $M^{'} = \alpha M + \beta I$, Thus,

$$\begin{aligned}(\alpha M + \beta I)x &= \alpha Mx + \beta Ix \\ &= \alpha(\lambda_i x) + \beta x \\ &= (\alpha\lambda_i + \beta)x\end{aligned}$$

Thus, the eigenvalue for $\alpha M + \beta I = (\alpha\lambda_i + \beta)$

2 Basic Statistics

2.1 Probabilities

Denote by A and B events, and by $\bar{A}$ the complement of A. Prove the following:

- Prove $\mathbb{P}(B \cap \bar{A}) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$

  Proof:
  We know that: $\mathbb{P}(A \cup B) = 1$

  Thus,
  $\mathbb{P}(B) = \mathbb{P}(\bar{A} \cap B) + \mathbb{P}(A \cap B)$

  We can substitute the above value of $\mathbb{P}(B)$ in the R.H.S of the equation to be proved

  $$\begin{aligned}R.H.S &= \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}((\bar{A} \cap B) + \mathbb{P}(A \cap B) - \mathbb{P}(A \cap B)) \\ &= P(\bar{A} \cap B) \\ &= L.H.S\end{aligned}$$

  Thus we have proved that,

  $$\mathbb{P}(B \cap \bar{A}) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

- Prove $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

  $$\begin{aligned}A &= A \cap \Omega \\ &= A \cap (B \cup \bar{B}) \\ &= (A \cap B) \cup (A \cap \bar{B}) \\ \Pr(A) &= \Pr(A \cap B) + \Pr(A \cap \bar{B}) \\ A \cup B &= (A \cup B) \cap (B \cup \bar{B}) \\ &= (A \cap \bar{B}) \cup B \\ &= (A \cap \bar{B}) \cup (B \cap (A \cup \bar{A})) \\ &= (A \cap \bar{B}) \cup (B \cap \bar{A}) \cup (A \cap B) \\ \Pr(A \cup B) &= \Pr(A \cap \bar{B}) + \Pr(A \cap \bar{B}) + \Pr(A \cap B) \\ &= \Pr(A) - \Pr(A \cap B) + \Pr(B) - \Pr(A \cap B) + \Pr(A \cap B) \\ &= \Pr(A) + \Pr(B) - \Pr(A \cap B)\end{aligned}$$

- Prove If $A \subset B$ then $\Pr(A) \leq \Pr(B)$

$$\begin{aligned} P(B) &= P(A \cup (\bar{A} \cap B)) \\ &= P(A) + P(\bar{A} \cap B) \\ &\geq P(A) + 0 \\ &\geq P(A) \end{aligned}$$

Hence proved,

$$P(A) \geq P(B)$$

2.2 Gaussian Distribution

The Gaussian distribution of parameters μ and σ^2 is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Prove that the mean and variance are μ and σ^2 respectively

Proof: The mean of a distribution is same as the expected value. We know that the $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ Thus, for the Gaussian Distribution

$$E[X] = \int_{-\infty}^{\infty} x\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx$$

We replace $x = y + \mu$, thus we have the following equation

$$E[X] = \int_{-\infty}^{\infty} (y+\mu)\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(y)^2}{2\sigma^2}\right\}dy$$

We can split the integral into 2 parts as below

$$E[X] = \int_{-\infty}^{\infty} y\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy \ + \int_{-\infty}^{\infty} \mu\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy$$

Let I_1 and I_2 be the first and the second parts of the above equation respectively. We can split the limits of I_1 into 2 parts and interchange the limits for the first part by adding a negative sign. Thus,

$$I_1 = -\int_{0}^{-\infty} y\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy \ + \int_{0}^{\infty} y\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy$$

We can replace y with $-y$ and and change the $-\infty$ to ∞ as it would be the same area under the curve. Thus we get,

$$I_1 = \int_{0}^{\infty} -y\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(-y)^2}{2\sigma^2}\right\}dy \ + \int_{0}^{\infty} y\frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy$$

$$I_1 = -\int_0^{\infty} y\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy \; + \int_0^{\infty} y\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy$$

Thus,

$$I_1 = 0$$

Now, let's look at I_2

$$I_2 = \int_{-\infty}^{\infty} \mu\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{y^2}{2\sigma^2}\right\}dy$$

We can substitute $t = \frac{y}{\sqrt{2}\sigma}$ and thus we would get,

$$I_2 = \int_{-\infty}^{\infty} \mu\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(\sqrt{2}\sigma t)^2}{2\sigma^2}\right\}dt$$

$$I_2 = \mu\left(\frac{2}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-t^2}dt\right)$$

Converting this to limits, the term in the bracket sums to 1. Thus we have,

$$I_2 = 1$$

Hence proved,

$$E[X] = \mu$$

Now, we have to prove that the

$$Var[X] = \sigma^2$$

<u>Proof</u>: We know that,

$$Var[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

Substituting the values for $f(x)$ we get,

$$Var[X] = \int_{-\infty}^{\infty} (x-\mu)^2\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx$$

Substituting $y = \sqrt{2}\sigma(x-\mu)$, we get,

$$Var[X] = \sqrt{2}\sigma\int_{-\infty}^{\infty} (\sqrt{2}\sigma y)^2\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(\sqrt{2}\sigma y)^2}{2\sigma^2}\right\}dy$$

Simplifying the above equation, we get

$$Var[X] = \frac{4\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} y^2 e^{y^2}dy$$

Substituting $t = y^2, t = \sqrt{y}$ *and* $dt = 2ydy$, we get

$$Var[X] = \frac{4\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} te^{-t} \frac{dt}{2\sqrt{t}}$$

$$Var[X] = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} \sqrt{t}e^{-t} dt$$

$$Var[X] = \frac{4\sigma^2}{\sqrt{\pi}} \frac{1}{2} \int_{-\infty}^{\infty} t^{\frac{3}{2}-1} e^{-t} dt$$

2.3 Poisson Distribution

The Poisson distribution has one parameter, the average rate $\lambda > 0$ and has probability mass function as below

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

We know that,

$$\sum Pr(X) = 1$$

The mean of Poisson distribution would be as below

$$E[X] = \sum_{k=0}^{\infty} kf(k)dx$$

Substituting $f(x)$ with the probability mass function, we get

$$E[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} dx$$

We can rewrite this as,

$$E[X] = \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} dx$$

$$E[X] = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} dx$$

In the above equation, the term inside the summation is the total probability of all the terms from 1 to ∞ and hence they sum to 1 Thus,

$$E[X] = \lambda$$

We know that $Var[X] = E[X^2] - (E[X])^2$, and we know that $E[X] = \lambda$. So let's first calculate $E[X(X-1)]$

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E[X(X-1)] = \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2} e^{-\lambda}}{(k-2)!}$$

Again, the term under the summation is the total probability and is thus equal to 1. Thus,

$$E[X(X-1)] = \lambda^2$$

$$E[X^2 - X] = \lambda^2$$

By linearity of expectation,

$$E[X^2] - E[X] = \lambda^2$$

And since, $E[X] = \lambda$

$$E[X^2] = \lambda^2 + \lambda$$

Substituting the value of $E[X^2]$ in the equation for variance, we get

$$Var[X] = \lambda^2 + \lambda - \lambda^2$$

Thus,

$$Var[X] = \lambda$$

Let $X_1 = Poisson(\lambda_1)$ and $X2 = Poisson(\lambda_2)$ be independent random variables. Show that the random variable $Z = X_1 + X_2$ is Poisson-distributed and compute its mean.

$$\begin{aligned} Pr(X_1 + X_2 = k) &= \sum_{i=0}^{k} \Pr(X_1 + X_2 = k, X_1 = i) \\ &= \sum_{i=0}^{k} \Pr(X_2 = k - i, X_1 = i) \end{aligned}$$

Since, X_1 and X_2 are independent

$$\begin{aligned}
Pr(X_1 + X_2 = k) &= \sum_{i=0}^{k} \Pr(X_2 = k - i) \Pr(X_1 = i) \\
&= \sum_{i=0}^{k} e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_1} \frac{\lambda_1^i}{i!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{k} \frac{\lambda_2^{k-i} \lambda_1^i}{i!(k-i)!} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \lambda_2^{k-i} \lambda_1^i \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda_2^{k-i} \lambda_1^i
\end{aligned}$$

The summation term is Binomial expansion

$$\begin{aligned}
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} (\lambda_1 + \lambda_2)^k \\
&= \frac{(\lambda_1 + \lambda_2)^k}{k!} \cdot e^{-(\lambda_1+\lambda_2)} \\
&= Poisson(\lambda_1 + \lambda_2)
\end{aligned}$$

We know that, the mean of Poisson(λ) is $\lambda. Thus,$

$$\begin{aligned}
E[Z] &= E[Poisson(\lambda_1 + \lambda_2)] \\
&= \lambda_1 + \lambda_2
\end{aligned}$$

$\blacksquare$

2.4 Estimators

1. Let $X_1, ..., X_n$ be i.i.d. random variables with mean μ and variance σ^2 M_n is a random variable such that

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$S_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_n)^2$$

Prove that $E[M_n] = \mu$ and $E[S_n] = \sigma^2$

Proof:

$$E[M_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

By linearity of expectation

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

Since X_i are i.i.d with a mean ofμ

$$= \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \frac{1}{n} n\mu$$

$$= \mu$$

Now let's prove $E[S_n] = \sigma^2$

$$E\left[S_n\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i^2 + M_n^2 - 2X_iM_n)\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} M_n^2 - \sum_{i=1}^{n} 2X_iM_n\right]$$

Since M_n is independent of X_i

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 + M_n^2\sum_{i=1}^{n} 1 - 2M_n\sum_{i=1}^{n} X_i\right]$$

Since $\sum X_i = M_n$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 + nM_n^2 - 2nM_n^2\right]$$

$$= \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n} X_i^2\right] - E\left[nM_n^2\right]\right)$$

$$= \frac{1}{n-1}\left(nE\left[X_i^2\right] - nE\left[M_n^2\right]\right)$$

We know that

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

Thus,

$$E[X_i^2] = \sigma^2 + \mu^2$$

Also we know that,

$$
\begin{aligned}
Var\left[M_n\right] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] \\
E\left[M_n^2\right] - E\left[M_n\right]^2 &= \frac{\sigma^2}{n} \\
E\left[M_n\right]^2 &= \frac{\sigma^2}{n} + \mu^2
\end{aligned}
$$

Thus from the above equations, we can have

$$
\begin{aligned}
E\left[S_n\right] &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\right) \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
E\left[S_n\right] &= \sigma^2
\end{aligned}
$$

Hence proved.

0.5 Ω_.png Ω_.pdf Ω_.jpg Ω_.mps Ω_.jpeg Ω_.jbig2 Ω_.jb2 Ω_.PNG Ω_.PDF Ω_.JPG Ω_.JPEG Ω_.JBIG2 Ω_.JB2 Ω_.eps

Figure 1: Convergence of M_n for $n \in [1, 10]$

0.5 .png .pdf .jpg .mps .jpeg .jbig2 .jb2 .PNG .PDF .JPG .JPEG .JBIG2 .JB2 .eps

Figure 2: Convergence of S_n for $n \in [1, 10]$

Figure 3: Testcase: $n \in [1, 10]$

2.

3 Agnostic PAC learning

4 Least Square Regression

4.1 Question 8

1. **trainls** Below is the function to find w, w_0 depending on rank of the input matrix and the bias

```
function [w, w_0] = train_ls(X, y, bias)
[m,n] = size(X);
```

```
    if bias==1
        disp('The bias is 1')
        Z = ones(m,1) ;
        X = [Z X]
    end
    if rank(X) ~= min(size(X))
        disp('Matrix is not full rank')
        [U,S,V] = svd(transpose(X)*X)
        Dplus = spfun(@(x) x.^-1, S)
        res = U*Dplus*transpose(U) * transpose(X)* y
    else
        disp('Matrix is full rank')
        res = inv(transpose(X)*X) * (transpose(X) * y);
    end
    if bias == 1
        w = res(2:n)
        w_0 = res(1)
    else
        w = res
        w_0 = 0
    end
end
```

2. **incrementaltrainls** The below code uses the Sherman-Morrison formula to update the inverse of the matrix $X^T X$

```
function [ w ] = incremental_train_ls( Xtrain, ytrain )
%   This function will call incremental_ls with a single row on the input %   and outpu
    global is_first;
    is_first = 1;
    m = size(Xtrain);
    w = zeros(1);
    for i = 1:m(1)
%         disp(Xtrain(i:i+1))
%         disp(ytrain(i:i+1))
        w = incremental_ls(Xtrain(i,:), ytrain(i,:));
    end
end
```

incrementalls This function is responsible for calculating w based of the input.

```
function [w] =  incremental_ls(X, y)
%     This is the incremental X matrix
    global A;
    global Y
```

```
    global Ainv;
    global is_first;
%   In the first iteration we calculate the inverse using standard methods
%   as Ainv will be empty initially. After the first iteration we update
%   the inverse by calling the sherman-morrison method
    if is_first == 1
        A = X
        Y = y
        X = transpose(X)*X;
        Ainv = inv(X);
        is_first = 0;
    else
        A = [A; X];
        Y = [Y; y];
        Ainv = woodburg_inverse(X, Ainv);
    end
    w = Ainv * (transpose(A)*Y)
end
```

woodburginverse This method updates the inverse *Ainv* using the Sherman Morrison formula

```
function [Ainv] = woodburg_inverse(X,Ainv)
    Ainv = Ainv - ((Ainv * (X * transpose(X))* Ainv)/(1+(X*Ainv*transpose(X))));
end
```

3. The solutions of the 2 algorithms on a random training set returns the same results, except in the case, when the matrix A in the second method is not full rank and thus the calculating the inverse is not possible as the matrix is not invertible.

4. **Complexity** The first method calculates the inverse of $X^T X$ in each iteration which takes $O(n^3)$ time. The Sherman-Morrison method is better in terms of time complexity because in each iteration, we are doing only $O(n^2)$

4.2 Question 9

References