Statistical Learning HW 1

Johnson Lin

October 14, 2019

1 Problem 1

Let Z $N(0, \sigma^2)$. *Show that* $sup_{t>0}\{P(Z \geq t)e^{t^2/(2\sigma^2)} = \frac{1}{2}\}$

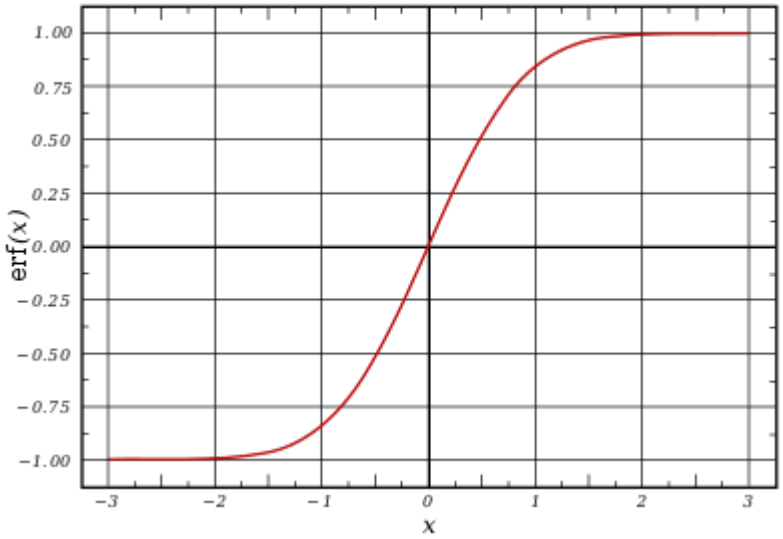
$P(Z \geq t)$ is the cdf of a normal function which can be defined as

$$P(Z \geq t) = 1 - \frac{1}{2}[1 + \text{erf}(\frac{x - \mu}{\sigma\sqrt{2}})]$$

In this case, x is t and $/mu = 0$, so

$$P(Z \geq t) = 1 - \frac{1}{2}[1 + \text{erf}(\frac{t}{\sigma\sqrt{2}})]$$

Looking at the graph of an erf function, we can see that when $\text{erf}(0) = \frac{1}{2}$ and as $x \to \infty$, $\text{erf}(x) \to 1$:



Because we are looking for the maximum when $t > 0$, the maximum must be when $t = 0 or \infty$. If $t = \infty$, then

$$P(Z \geq t) = 1 - \frac{1}{2}[1+1]$$

$$P(Z \geq t) = 1 - \frac{1}{2}[2]$$

$$P(Z \geq t) = 1 - 1$$

$$P(Z \geq t) = 0$$

Thus the overall equation equals 0. So our maximum must be when $t = 0$:

$$P(Z \geq t) = 1 - \frac{1}{2}[1+0]$$

$$P(Z \geq t) = 1 - \frac{1}{2}[1]$$

$$P(Z \geq t) = 1 - \frac{1}{2}$$

$$P(Z \geq t) = \frac{1}{2}$$

At $t = 0$, $e^{t^2/(2\sigma^2)} = e^0 = 1$. Thus, our equation becomes:

$$\frac{1}{2} * 1$$

Thus proving

$$sup_{t>0}\{P(Z \geq t)e^{t^2/(2\sigma^2)} = \frac{1}{2}\}$$

2 Problem 2

Consider the covariance matrix $\Sigma = (\sigma_{ij})$ with an autoregressive Toeplitz structure: $\sigma_{ij} = \rho^{|i-j|}$ with $0 < |\rho| < 1$. Show that the irrepresentable condition holds and identify the constant a

Our covariance matrix is a $n \times n$ matrix that looks like this:

$$\Sigma = \begin{bmatrix} 1 & \rho & \dots & \rho^{j-1} \\ \rho & 1 & \dots & \rho^{j-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{i-1} & \rho^{i-2} & \dots & 1 \end{bmatrix}$$

The eigenvalues for this matrix are $e_1 = 1 + (n-1)\rho^{|i-j|}$ and $e_i = 1 - \rho^{|i-j|}$ for $i \geq 2$. Thus the inverse of this matrix is

$$(\Sigma)^{-1} = \begin{bmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{bmatrix}$$

And its eigenvalues are $e_1' = c + (q-1)d = \frac{1}{e_1} = \frac{1}{1+(n-1)\rho^{|i-j|}}$.

We can define $\Sigma_{S^C S} = \rho^{|i-j|} \times \mathbf{1}_{(p-n)\times n} = \frac{\rho^{|i-j|}}{1+(n-1)\rho^{|i-j|}} \mathbf{1}_{(p-n)\times n}$

So,

$$|\Sigma_{S^c S}(\Sigma_{SS})^{-1} sign(\beta_S^*)| = \frac{\rho^{|i-j|}}{1+(n-1)\rho^{|i-j|}} |\Sigma sgn\beta_i| \mathbf{1}_{(n\times 1}$$

$$\leq \frac{n\rho^{|i-j|}}{1+(n-1)\rho^{|i-j|}} \mathbf{1}_{(n\times 1} \leq \frac{\frac{n}{1+cn}}{1+\frac{n-1}{1+cn}} = \frac{1}{1+c}$$

Thus, the Strong Irrepresentable Condition holds. In this case, the constant $a = \frac{1}{1+c}$

3 Problem 3

Derive the ADMM algorithm for the group Lasso problem

ADMM consists of these iterations:

$$x^{k+1} := \text{argmin}_x L_\rho(x, z^k, y^k)$$

$$z^{k+1} := \text{argmin}_z L_\rho(x^{k+1}, z, y^k)$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

In terms of lasso, we can think of it as minimizing $f(x) + g(z)$ where $f(x) = \frac{1}{2}||Ax-b||_2^2$ and $g(z) = \lambda||z||_1$.

So the ADMM algorithm for the lasso problem consists of these iterations:

$$x^{k+1} := (A^T A + \rho I)^{-1}(A^T b + \rho(z^k - u^k))$$

$$z^{k+1} := S_{\lambda/\rho}(x^{k+1} + u^k)$$

$$u^{k+1} := u^k + x^{k+1} - z^{k+1}$$

In Group Lasso, instead of the regularizer $||x||_1$, we replace it with $\Sigma_{i=1}^N ||x_i||_2$. So our new objective is to minimize $\frac{1}{2}||Ax-b||_2^2 + \lambda\Sigma_{i=1}^N ||x_i||_2$ This change affects the update of the z variable which becomes:

$$z^{k+1} := S_{\lambda/\rho}(x_i^{k+1} + u^k)$$

for $i = 1, \ldots, N$ where$S_\kappa(a) = (1 - \kappa/||a||_2)_+ a$

The rest of the iterations remain the same.

4 Problem 4

ESL Ex. 3.30

Augment X with a multiple of the $p \times p$ identity and augment y with p zero values:

$$X = \begin{bmatrix} X \\ \gamma I \end{bmatrix} y = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

This turns our problem into

$$||y - X\beta||_2^2 = || \begin{bmatrix} y - X\beta \\ -\gamma\beta \end{bmatrix} ||_2^2$$

Because we are squaring and taking the absolute value, the negative sign doesn't matter

$$= || \begin{bmatrix} y - X\beta \\ \gamma\beta \end{bmatrix} ||_2^2 = ||y - X\beta||_2^2 + \gamma^2||\beta||_2^2$$

The lasso problem becomes

$$\hat{\beta} = \text{argmin}_\beta(||y - X\beta||_2^2 + \gamma^2||\beta||_2^2 + \lambda||\beta||_1)$$

If we set $\gamma^2 = \lambda a$ and $\lambda = \lambda(1 - a)$, then we get the original problem:

$$\hat{\beta} = \text{argmin}_\beta(||y - X\beta||_2^2 + a\lambda||\beta||_2^2 + (1 - a)\lambda||\beta||_1)$$

$$\hat{\beta} = \text{argmin}_\beta(||y - X\beta||_2^2 + \lambda(a||\beta||_2^2 + (1 - a)||\beta||_1))$$

By augmenting X and y, we have effectively turned it into a lasso problem.

5 Problem 5

ESL Ex. 4.2

a. Bayes' discriminant function is

$$\delta_k(x) = ln(p(x|\omega_k)) + ln(\pi_k)$$

If we let our conditional density $p(x|\omega_k)$ be given by a normal distribution, then

$$p(x|\omega_k) = N(x; \mu_k; \Sigma_k) \equiv \frac{1}{(2\pi)^{p/2}|\Sigma_k|^(1/2)} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

$$ln(p(x|\omega_k)) = -\frac{1}{2}(x - \mu_k)^T\Sigma_k^(- 1) - \frac{p}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_k|)$$

Plugging this in to our original equation, we get the discriminant function

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{p}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_k|) + ln(\pi_k)$$

We can expand this

$$\delta_k(x) = -\frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k) - \frac{p}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_k|) + ln(\pi_k)$$

$$\delta_k(x) = -\frac{1}{2}x\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_k + \frac{1}{2}\mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{p}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_k|) + ln(\pi_k)$$

Because $x^T\Sigma^{-1}\mu_k$ is a scalar and Σ^{-1} is symmetric,

$$x^T\Sigma^{-1}\mu_k = (x^T\Sigma^{-1}\mu_k)^T = \mu_k^T(\Sigma^{-1})^T x = \mu_k^T\Sigma^{-1}x$$

We can now combine and simplify our equation

$$\delta_k(x) = -\frac{1}{2}x\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{p}{2}ln(2\pi) - \frac{1}{2}ln(|\Sigma_k|) + ln(\pi_k)$$

We can drop some of the values because we are looking at the specialization of this expression to get

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + ln(\pi_k)$$

Now if we estimate $\pi_i = \frac{N_i}{N}$ for $i = 1, 2$ and classify it as $\delta_2(x) > \delta_1(x)$ and class 1 otherwise. This creats the inequality

$$x^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + ln(\frac{N_2}{N}) > x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + ln(\frac{N_1}{N})$$

$$x^T\Sigma^{-1}(\mu_2 - \mu_2) > \frac{1}{2}\mu_1^T\sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + ln(\frac{N_1}{N}) - ln(\frac{N_2}{N})$$

b. We are looking to minimize $\Sigma_{i=1}^N(y_i - \beta_0 - \beta^T x_i)^2$ and the solution must satisfy the normal equation so

$$X^T X \begin{bmatrix}\beta_0 \\ \beta\end{bmatrix} = X^T y$$

$$X^T X = \begin{bmatrix} N & \sigma_{i=1}^N x_i^T \\ \sigma_{i=1}^N x_i & \sigma_{i=1}^N x_i x_i^T \end{bmatrix}$$

$$= \begin{bmatrix} N & N_1\mu_1^T + N_2\mu_2^T \\ N_1\mu_1 + N_2\mu_2 & (N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T \end{bmatrix}$$

In this case, our response is coded as $-\frac{N}{N_1}$ for the first class and $+\frac{N}{N_2}$ for the second class, where $N = N_1 + N_2$, $X^T y$ becomes

$$\left[N_1(-\tfrac{N}{N_1}) + N_2(\tfrac{N}{N_2})//(\Sigma_{i=1}^{N_1}x_i)(-\tfrac{N}{N_1}) + (\Sigma_{i=N_1+1}^{N}x_i)(\tfrac{N}{N_2})\right] = \begin{bmatrix} 0 \\ -N\mu_1 + N\mu_2 \end{bmatrix}$$

So our normal equation becomes

$$\begin{bmatrix} N & N_1\mu_1^T + N_2\mu_2^T \\ N_1\mu_1 + N_2\mu_2 & (N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ -N\mu_1 + N\mu_2 \end{bmatrix}$$

The first equation in the system is

$$N\beta_0 + (N_1\mu_1^T + N_2\mu_2^T)\beta = 0$$

$$\beta_0 = (\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T)\beta$$

Plugging this back in, we get

$$N((\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T)\beta) + (N_1\mu_1^T + N_2\mu_2^T)\beta = 0$$

$$(N_1\mu_1+N_2\mu_2)(-\frac{N_1}{N}\mu_1^T-\frac{N_2}{N}\mu_2^T)\beta+((N-2)\hat{\Sigma}+N_1\mu_1\mu_1^T+N_2\mu_2\mu_2^T)\beta = N(\mu_2-\mu_1)$$

$$[(N_1\mu_1+N_2\mu_2)(-\frac{N_1}{N}\mu_1^T-\frac{N_2}{N}\mu_2^T)((N-2)\hat{\Sigma}+N_1\mu_1\mu_1^T+N_2\mu_2\mu_2^T)]\beta = N(\mu_2-\mu_1)$$

Let's focus on the outer product terms

$$= -\frac{N_1^2}{N}\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T - \frac{N_2^2}{N}\mu_1\mu_2^T + N_1\mu_1\mu_2^T + N_2\mu_2\mu_2^T$$

$$= (-\frac{N_1^2}{N} + N_1)\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + (-\frac{N_2^2}{N} + N_2)\mu_2\mu_2^T$$

$$= \frac{N_1}{N}(-N+1+N)\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \frac{N_2}{N}(-N_2+N)\mu_2\mu_2^T$$

$$= \frac{N_1N_2}{N}\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \frac{N_2N_1}{N}\mu_2\mu_2^T$$

$$= \frac{N_1N_2}{N}(\mu_1\mu_1^T - 2\mu_1\mu_2^T + \mu_2\mu_2^T)$$

$$= \frac{N_1N_2}{N}(\mu_1-\mu_2)(\mu_1-\mu_2)^T$$

We know that

$$\hat{\Sigma}_B \equiv (\mu_1-\mu_2)(\mu_1-\mu_2)^T$$

so we get the equation we wanted to show

$$[(N-2)\hat{\Sigma} + \frac{N_1N_2}{N}\hat{\Sigma}_B]\beta = N(\mu_2-\mu_1)$$

c.

$$\hat{\Sigma}_B\beta = (\mu_2-\mu_1)(\mu_2-\mu 1)^T\beta$$

Because $(\mu_2 - \mu 1)^T\beta$ is a scalar due to both μ_i and β being $n \times 1$ matrices. The product of the transpose and β gives a 1×1 matrix which is a scalar.

Thus since it is a scalar, the direction must be dependent on the direction of $(\mu_2 - \mu_1)$. Hence β must be in the same direction and therefore proportional to $\hat{\Sigma}^{-1}((\mu_2 - \mu_1))$

d. The result of (c) isn't dependent on the coding of the two classes. The equation holds true for any coding

$$\hat{\Sigma}_B \beta = (\mu_2 - \mu_1)(\mu_2 - \mu 1)^T \beta$$

The product of $(\mu_2 - \mu 1)^T \beta$ will always be scalar because they will always be the same sized matrix. Thus the result holds for any coding of the two classes.

e. We got what β_0 is from above

$$\beta_0 = (\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T)\beta$$

So our predicted value $\hat{f}(x)$ is

$$\hat{f}(x) = (\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T)\beta + \beta^T x$$

$$\hat{f}(x) = (\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T)\lambda(\Sigma)^{-1}(\mu_2 - \mu_1) + \beta^T x$$

Our classification rule is $y_i > 0$ so we can drop some of the common terms to get

$$Nx^T\Sigma^{-1}(\mu_2 - \mu_1) > (N_1\mu_1^T + N_2\mu_2^T)\Sigma^{-1}(\mu_2 - \mu 1)$$

$$x^T\Sigma^{-1}(\mu_2 - \mu_1) > \frac{1}{N}(N_1\mu_1^T + N_2\mu_2^T)\Sigma^{-1}(\mu_2 - \mu 1)$$

which is only equivalent to LDA when $N_1 = N_2$.

6 Problem 6

ESL Ex. 4.3

The discriminant function for $\hat{Y}$ would be

$$\delta_k = \hat{Y}\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + log\pi_k$$

If we expand the first expression in the equation we get

$$\hat{Y}\hat{\Sigma}^{-1}\hat{\mu}_k = (X\beta)(\beta^T\Sigma\beta)^{-1}(\frac{\beta^T X^T Y_k}{N_k})$$

$$\hat{Y}\hat{\Sigma}^{-1}\hat{\mu} = (X\beta)(\beta^T\Sigma\beta)^{-1}(\beta^T X^T Y D^{-1})$$
$$= (X\beta)(\beta^{-1}\Sigma^{-1}(\beta^T)^{-1})(\beta^T X^T Y D^{-1})$$
$$= X\Sigma^{-1}X^T Y D^{-1}$$
$$= X\Sigma^{-1}\mu$$

If we expand the second expression in the equation we get

$$\begin{aligned}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k = \hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu} &= (\beta^T\mu_k)^T\hat{\Sigma}^{-1}(\beta^T X^T Y D^{-1}) \\ &= \mu_k^T\beta\hat{\Sigma}^{-1}(\beta^T X^T Y D^{-1}) \\ &= \mu_k^T\beta(\beta^T\Sigma\beta)^{-1}(\beta^T X^T Y D^{-1}) \\ &= \mu_k^T\beta\beta^{-1}\Sigma^{-1}(\beta^T)^{-1}(\beta^T X^T Y D^{-1}) \\ &= \mu_k^T\Sigma^{-1}X^T Y D^{-1} \\ &= \mu_k^T\Sigma^{-1}\mu\end{aligned}$$

If we put it all together,

$$\delta_k = X\Sigma^{-1}\mu - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu + log\pi_k$$

this is the original discriminant function which shows that LDA on $\hat{Y}$ is equivalent to LDA on X.

7 Problem 7

ESL Ex. 4.5

The log-likelihood function is

$$\ell(\beta) = \Sigma_{i=1}^N(y_i\beta^T x_i - log(1 + e^{\beta^T x_i})$$

Since the data is separable around x_0, we can define it as

$$y_i = \begin{cases} 0 & x_i \leq x_0 \\ 1 & x_i > x_0 \end{cases}$$

So we can write the log-likelihood function as

$$\ell(\beta) = \Sigma_{i=1,x_i\leq x_0}^N - log(1 + e^{\beta_0+\beta_1 x_i} + \Sigma_{i=1,x_i>x_0}^N \beta_0 + \beta_1 x_i - log(1 + e^{\beta_0+\beta_1 x_i}$$

$$= \Sigma_{i=1,x_i\leq x_0}^N - log(1+e^{\beta_0+\beta_1 x_0+\beta_1(x_i-x_0)} + \Sigma_{i=1,x_i>x_0}^N \beta_0+\beta_1 x_0+\beta_1(x_i-x_0) - log(1+e^{\beta_0+\beta_1 x_0+\beta_1(x_i-x_0)}$$

Our goal is to maximize this function. If we set $\beta_0 = -\beta 1 x_0$, then our equation becomes

$$\ell(\beta) = \Sigma_{i=1,x_i\leq x_0}^N - log(1+e^{\beta_1(x_i-x_0)} + \Sigma_{i=1,x_i>x_0}^N \beta_1(x_i-x_0) - log(1+e^{\beta_1(x_i-x_0)}$$

In this expression, in order to maximize it, we should let $\beta_i \to \infty$. The first part of the expression is maximized the smaller the value of the expression inside *log*. Because $x_0 \geq x_i$, by increasing β_1 we are decreasing the value of the expression. The second part of the expression is also maximized by letting $\beta_i \to \infty$ since $(x_i - x_0)$ is positive in this expression. So the solution for this problem is by letting $\beta_1 \to \infty$ and $\beta_0 \to -sign(x_0)\infty$

For two classes, this is separable by a single hyperplane. Hence, for multiple classes, it is separable by $K - 1$ hyerplanes where K is the number of classes.

8 Problem 8

ESL Ex. 4.7

By setting the constraint $||\beta|| = 1$, the expression $x_i^T\beta + \beta_0$ becomes the distance from the point x_i to the hyperplane given by $x^T\beta + \beta_0 = 0$. So when $D(\beta, \beta_0)$ is minimized, x_i will have a high positive value when $y_i = 1$ and a negative value when $y_i = -1$. This is how it finds the separating hyperplane between the two classes. It's not perfect however because it is only taking into account the sum of the distance so it can be susceptible to outliers. It does not solve the optimal separating hyperplane problem.