CS 189: Homework 2

William Guss
26793499
wguss@berkeley.edu

February 19, 2016

1. Let $a = 1/\sqrt{3}, b = 1, c = \sqrt{3}$. Then recall that

$$E[G] = 4\int_0^a f(x)dx + 3\int_a^b f(x)dx + 2\int_b^c f(x)dx. \tag{1}$$

Using babies first calculus class, we get that

$$\int f(x)dx = \frac{\arctan(x)}{\pi}, \tag{2}$$

giving us

$$E[G] = \frac{4}{3} + \frac{1}{3} + \frac{1}{3} = 2. \tag{3}$$

2. Maximum Likelihood Estimation? Recall from wikipedia, that since $f(x;\theta)$ for each is generated independently and identically distributed, we have

$$f(x_1, x_2 \ldots; \theta) = \prod_{i}^{n} \theta e^{-\theta x_n}. \tag{4}$$

We want to find a value θ which maximizes the average log-likihood, given by

$$\ell = \frac{1}{n} \sum \ln(\theta e^{-\theta x_i}) = \sum \ln(\theta) - \theta x_i. \tag{5}$$

Using calculus, we look for θ satisfying

$$\begin{aligned} \ell' = \sum \frac{1}{\theta} - x_i = 0 \\ \frac{n}{\theta} = \sum x_i \\ \theta = \frac{n}{\sum x_i}. \end{aligned} \tag{6}$$

Applying the values we get $\sum x_i = 5.9$, $n = 5$, and $\theta = 0.847457627$.

3. Let A be a positive definite matrix in $\mathbb{R}^{n\times n}$.

   (a) Consider the following derrivation:

$$x^T A x = x^T \begin{bmatrix} \sum_j^n a_{1j}x_j \\ \vdots \\ \sum_j^n a_{nj}x_j \end{bmatrix} = \sum_i^n \sum_j^n a_{ij}x_i x_j. \tag{7}$$

   (b)

Theorem 1. *If A is positive definite, then the diagonals of A are positive.*

Proof. Suppose that there is negative value on the diagonal, say a_{qq}. Then let $x = e_q$. If we apply the quadratic form we get $e_q A^T e_q = a_q q < 0$. This contradicts the positive semidefiniteness of A. $\square$

4. Short Proofs.

   (a) Assume problem (b).

   Lemma 1. *If A is a matrix with eigen values λ_n $A+\gamma I$ has eigenvalues $\gamma+\lambda_n$*

   Proof. If λ_n is an eigenvalue, then $Av_n = \lambda_n v_n$ for a corresponding eigenvector v. Furthermore

   $$(A+\gamma I)v = Av + \gamma Iv = \lambda_n v + \gamma v = (\lambda_n + \gamma) \tag{8}$$

   which implies that $\lambda_n + \gamma$ is an eigen value of $A+\gamma I$. This completes the proof. □

   Theorem 2. *If A is positive semidefinite and $\gamma > 0$, then $A+\gamma I$ is positive definite.*

   Proof. If A is positive definite then by the logic of the proof of (b),

   $$x^T Ax = \sum_i \lambda_i (x_i^T e_i)^2 \geq 0. \tag{9}$$

   It follows that some $\lambda \geq 0$ since $x \neq 0$. Therefore by the previous lemma adding γ to the diagonal adds γ to every eigenvalue implying that all eigen values are positive. By (b), $A + I\gamma$ is positive definiute therefore. □

   (b) Lolololol!

   Theorem 3. *A is positive definite if and only if all of its eigen values are more than 0.*

   Proof. Iff A is positive semidefinite then it is symmetric. Using spectral theorem we have that

   $$\begin{aligned} x^T Ax = \sum_i (x^T e_i) e_i^T Ax = \sum_i &= x^T e_i e_i^T \lambda e_i^T x \\ &= \sum_i \lambda_i (x^T e_i)^2 > 0 \end{aligned} \tag{10}$$

   which is true if and only if all λ_i are more than 0. □

   (c)

   Theorem 4. *If A is positive definite then it is invertible.*

   Proof. The invertible matrix theorem statesa that a matrix is invertible if and only if all of its eigen values are more than 0. By the previous theorem if A is positive definite then all of its eigen values are positive and so it is invertible. □

   (d)

   Theorem 5. *If A is positive definite then there exist n linearly independent vectors so that $A_{ij} = x_i^T x_j$.*

Proof. The statement of the theorem is true if and only if $A = B^T B$ where B is invertible. By spectral theorem we have that $A = U\Lambda U^T$ where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$. Furthermore $U^{-1} = U^T$. Let $\Omega = diag(\sqrt{\lambda}_1, \ldots, \sqrt{\lambda}_n)$. Then, $\Omega^2 = \Lambda$. Let $W^T = U\Omega$ and $W = \Omega U^T$. So we have that W is still an orthonormal matrix and so $A = W^T W$. This completes the proof. $\square$

5. DERIVATIONS :(Assuming theorems from Math 105

(a) Consider the followiong derivation

$$\frac{\partial(x^T a)}{\partial x} = \frac{\partial(x)}{\partial x}^T a + \left(\frac{\partial(a)}{\partial x}\right)^T x = a. \tag{11}$$

(b) Consider the following derivation

$$\frac{\partial(x^T Ax)}{\partial x} = \frac{\partial(x^T)}{\partial x} Ax + \frac{\partial(Ax)}{\partial x}^T X = Ax + Ax^T \tag{12}$$

(c) Consider the following derivation

(d)

Theorem 6. *If* $x \in \mathbb{R}^n$

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{13}$$

Proof. Squaring the first two terms of the inequality shows that $\|x\|_2^2$ has fewer terms than $\|x\|_1$.

Now define the following vector, e, so that $e_i = 1$ if x_i is positive and $e_i = -1$ if x is negative. Then $\langle x, e\rangle = \sum_i |x_i| = \|x\|_1$.

Cauchy schwartz says that $\langle x, e\rangle \leq \|x\|\|e\| = \|x\|_2\sqrt{n}$. This completes the proof. $\square$

6. Weighted Linear Lolzs.

(a) Consider the following.

$$\begin{aligned} R[w] = \sum_i \lambda_i (w^T x_i - y_i)^2 &= \sum_i (w^T x_i - y_i) \lambda_i (w^T x_i - y_i) \\ &= \sum_i v^T \Lambda v. \end{aligned} \tag{14}$$

Observe that $v = (w^T x_i - y_i, \dots)^T = (w^T x_i, dots) - Y = Xw - Y$. Therefore

$$R[w] = (Xw - Y)^T \Lambda (Xw - Y). \tag{15}$$

(b) We can use the linearity of matrix multiplication to derive the following expression:

$$\begin{aligned} \frac{\partial R}{\partial w} &= \frac{\partial}{\partial w}(Xw)^T \Lambda (Xw - Y) - Y^T \Lambda (Xw - Y) \\ &= \frac{\partial}{\partial w}(Xw)^T \Lambda X w - (Xw)^T \Lambda Y - Y^T \Lambda (Xw) + Y^T \Lambda Y \\ &= \frac{\partial}{\partial w}(Xw)^T \Lambda X w - (Xw)^T \Lambda Y - \frac{\partial}{\partial w} Y^T \Lambda X w \\ &= \frac{\partial}{\partial w}(Xw)^T \Lambda X w - (Xw)^T \Lambda Y - (Y^T \Lambda X) \\ &= X^T \Lambda X w + (\Lambda X)^T X w - X^T \Lambda Y - Y^T \Lambda X = 0. \end{aligned} \tag{16}$$

And so we can manipulate the expression so that

$$\begin{aligned} (X^T \Lambda X + (\Lambda X)^T X) w &= X^T \Lambda Y + Y^T \Lambda X \\ w &= ((X^T \Lambda X + (\Lambda X)^T X)^{-1} (X^T \Lambda Y + Y^T \Lambda X) \\ &= X^{-1} (X^T \Lambda + X^T \Lambda^T)^{-1} (X^T \Lambda Y + Y^T \Lambda X) \\ &= 2 X^{-1} (X^T \Lambda)^{-1} (X^T \Lambda Y + Y^T \Lambda X) \\ &= 2 (X^T \Lambda X)^{-1} (X^T \Lambda Y + Y^T \Lambda X) \end{aligned} \tag{17}$$

(c) Adding L_2 regularization! Gives us

$$R[w] = (Xw - Y)^T \Lambda (Xw - Y) + w^T \gamma I w. \tag{18}$$

Taking dthe derivative we get

$$\begin{aligned} \frac{\partial R}{\partial w} &= \frac{\partial}{\partial w}(Xw)^T \Lambda (Xw - Y) - Y^T \Lambda (Xw - Y) + \frac{\partial}{\partial w} w^T \gamma I w \\ &= \frac{\partial}{\partial w}(Xw)^T \Lambda X w - (Xw)^T \Lambda Y - Y^T \Lambda (Xw) + Y^T \Lambda Y + \frac{\partial}{\partial w} w^T \gamma I w \\ &= \frac{\partial}{\partial w}(Xw)^T \Lambda X w - (Xw)^T \Lambda Y - \frac{\partial}{\partial w} Y^T \Lambda X w + \frac{\partial}{\partial w} w^T \gamma I w \\ &= \frac{\partial}{\partial w}(Xw)^T \Lambda X w - (Xw)^T \Lambda Y - (Y^T \Lambda X) + \frac{\partial}{\partial w} w^T \gamma I w \\ &= X^T \Lambda X w + (\Lambda X)^T X w + 2\gamma I w - X^T \Lambda Y - Y^T \Lambda X = 0. \end{aligned} \tag{19}$$

And so we can manipulate the expression so that

$$\begin{aligned}
(X^T \Lambda X + (\Lambda X)^T X + 2I\gamma)w &= X^T \Lambda Y + Y^T \Lambda X \\
w &= ((X^T \Lambda X + (\Lambda X)^T X + 2I\gamma)^{-1}(X^T \Lambda Y + Y^T \Lambda X) \\
&= \frac{1}{2}(X^T \Lambda X + I\gamma)^{-1}(X^T \Lambda Y + Y^T \Lambda X).
\end{aligned} \tag{20}$$

Essentially we add to the least squares pseudo inverse γI, thereby increasing its eigen values. This may allow us to find a solution when $X^T \Lambda X$ is non-singular. ie. L_2 regulartization penalizes movement in any (infinite) direction too far away from a small solution, it also forces a solution with γ large enough.

7. Doubt Classes!

   (a) We wish to minimize risk with respect to i. So, observe the logic of the policy. If the probability that ω_j is the output given x is less than that with respect to i then we wish to eliminate this large contribution to the some. It must furthermore be that such a probability be at least less than the loss incurred by the doubt. That is consider the expected value $l(...) = \lambda_s$ in the case that we don't choose the doubt. So then $\lambda_s(1 - \lambda_r/\lambda_s) = \lambda_s - \lambda_r$ if and only if the doubt in this situation has less 'weight' than making the prediction. Therefore this policy makes sense.

   (b) In the case that there is no loss incurred by doubting, it follows that using the minimum risk strategy immediately implies that fior every training example we choose to doubt unless $P(\omega_i|x) = 1$, in which case we are 100% certain that our classification estimate is correct. This agrees with my intuition.

   In the case thgat $\lambda_r > \lambda_s$, the intuition is that choosing to doubt our predicition is more "negativeley" impactful then to choose our prediction itsself; that is, more loss is incurred if we tend towards a doubt class. Therefore by our "minimum" risk procedure, we should choose to accept the prediction instead of the doubt every single time. ($P(\omega_i|x) \geq 0 > 1 - m$ where $m > 1$.)

8. Gaussians

   (a) We want to equate the two distributions so:

$$\begin{aligned}
P(\omega_1|x) = \frac{1}{2}P(x|\omega_1) &= \frac{1}{2\sqrt{2\pi\sigma^2}}exp\left(\frac{(x-\mu_1)^2}{2\sigma^2}\right) \\
= \frac{1}{2}P(x|\omega_2) &= \frac{1}{2\sqrt{2\pi\sigma^2}}exp\left(\frac{(x-\mu_2)^2}{2\sigma^2}\right) \\
\implies (x-\mu_1)^2 &= (x-\mu_2)^2
\end{aligned} \tag{21}$$

   So we get the plus or minus definition of μ_1. This completes the derivation.