HOMEWORK 1

Bill Chambers

September 23, 2015

Contents

1 Problem 1 Solution 1

2 Problem 2 2

3 Problem 3 3

3.1 Part a 3

3.2 Part b 3

4 Problem 4 3

4.1 a 3

4.2 b 3

5 Problem 5 3

6 Problem 6 4

6.1 Part a 4

6.2 Part b 4

6.3 Part c 4

6.4 Part d 4

6.5 Part e 5

7 Problem 7 5

7.1 Part a 5

7.2 Part b 5

7.3 Part c 6

7.4 part d 6

8 Problem 8 6

8.1 Part a 6

8.2 Part b 6

1 Problem 1 Solution

CDF = $\int$ f(x) dx

$= \int \frac{2}{\pi(1+x^2)} dx$

$= \frac{2}{\pi} \int \frac{2}{1+x^2} dx$

= $\frac{2}{\pi} tan^{-1}$(x)

Now we just need to evaluate this for each interval. Keep in mind that each of these need to be multiplied by $2/\pi$.

tan^{-1}(0) = 0
tan^{-1}(1/$\sqrt{3}$) =$\pi/6$
tan^{-1}(1) = $\pi/4$
tan^{-1}($\sqrt{3}$) =$\pi/3$

We know that $\int_{ab}$ f(x)dx = F(b) - F(a). So we can use that to get our probabilities.

For the interval from 0 -> 1/$\sqrt{3}$
2/π * π/6 - 0 = 2/6

For the interval from 1/$\sqrt{3} - > 1$
2/π * π/4 - 2/π * π/6 = 1/6

For the interval from 1 -> $\sqrt{3}$
2/π * π/3 - 2/π * π/4 = 1/6

Now that we have those values we can go about calculating the expectation which is simply the probabilities of each of those happening (that we calculated above) multiplied by the point values in order to get the expected value.

4/6 + 3/6 + 2/6 = $\frac{13}{6}$

2 Problem 2

All sums/products are over N.

P(x_i | θ) = θ $e^{-\theta\ x}$
lik(θ) = $\prod^n$ p(x_i |θ)
= $\sum$ log(p(x_i |θ))
= $\sum$ log(θ) - θ x_i
= n log(θ) - $\sum$ θ x_i

Now we can set the derivative equal to 0 to get the maximum likelihood.
max. lik(θ) = ∂ y / ∂ θ = 0

= $\frac{\partial n log(\theta) - \sum \theta x_i}{\partial \theta}$ = 0
= n/θ - $\sum$ x_i = 0
= $\frac{n}{\theta}$ =$\sum$ x_i
= n/ $\sum$ x_i = θ
= $\frac{5}{5.7}$ =θ

3 Problem 3

3.1 Part a

3.2 Part b

Cross validation and leveraging that to tune hyperparameters. We want to get something that generalizes well and doesn't just minimize the empirical risk but rather the generalized risk.

4 Problem 4

4.1 a

i and j are equal to the rows and columns of the vectors respectively.
x^{T} A x = $\sum_{\mathrm{j}=1}^{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a}_{\mathrm{j}}^{\mathrm{i}} \, \mathrm{x}_{\mathrm{i}} \, \mathrm{x}_{\mathrm{j}}$

4.2 b

Let x be all vectors such that we have all vectors that have a 1 in one specific place and zero in all the others. For example,

$$\begin{bmatrix}1\\0\\0\end{bmatrix} \begin{bmatrix}0\\1\\0\end{bmatrix} \begin{bmatrix}0\\0\\1\end{bmatrix}$$

But continued for the length of N so that we can perform x^{T} A x. These vectors exist so that they can extract the diagonal values of matrix A. Given that x^{T} A x is necessarily greater than 0 for all vectors non-zero x, the diagonal elements have to be positive or else we've invalidated the fact that x^{T} A x > 0. Quite simply, the only way that x^{T} A x can be negative or zero, is if A diagonal values are negative or zero because even if x is negative, it will still result in a positive value because it's squared as we saw in the previous summation.

5 Problem 5

From the previous problem we proved that if B is a positive definite matrix, then all the diagonal values have to be positive. If we repeat that proof along with the assertion that for p.s.d. x^{T} Ax ≥ 0 we prove that all diagonal values in the p.s.d. matrix are greater than equal to 0.

Using that, we can see that if we extract the minimum value from the diagonal of the lowest p.s.d. matrix B, then we will extract a 0 value (and no less). Secondly if we extract the minimum value from the smallest possible matrix of γ I, that has to necessarily be positive if $\gamma > 0$. Therefore the minimum value from the diagonal from matrix B added to the minimum value from the diagonal of matrix γ I is necessarily positive. Since the minimum value has to be positive, then all the diagonals in B + γ I are necessarily positive.

Now we can apply the proof from the previous problem, that for all non-zero x, x^{T} Ax must be greater than 0. Our newly created matrix (B + γ I) satisfies this requirement because the only way that it would be ≤ 0 would be in one of the diagonal elements is 0 or negative. We just proved that our matrix B + γ I cannot have 0 or negative diagonal elements, therefore our matrix is now positive definite.

6 Problem 6

6.1 Part a

x^{T} a = $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}} * \mathrm{a}_{\mathrm{i}}$

$$\frac{\partial \sum_{i=1}^{n} x_i * a_i}{\partial [x_1....x_n]}$$

Derive it component by component and you get the vector a.

6.2 Part b

x^{T} A x = $\sum_{\mathrm{j}=1}^{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a}_{\mathrm{j}}^{\mathrm{i}} \mathrm{x}_{\mathrm{i}} \mathrm{x}_{\mathrm{j}}$

$$\frac{\partial \sum_{j=1}^{n} \sum_{i=1}^{n} a_j^i x_i x_j}{\partial [x_1....x_n]}$$

When you derive it component by component you end up with the derivatives of x on the columns added to the derivatives of x on the rows of matrix A (since x^{T} affects rows and x affects columns).

Ax + A^{T} x

6.3 Part c

$$\frac{\partial Trace(XA)}{\partial X}$$

$$Trace(XA) == \sum (XA)_{ii} == \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} a_{ji}$$

Now when we look at the derivative w.r.t. X. . . .

$$\frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} a_{ji}}{\partial \begin{bmatrix} x_{1,1} & ... & x_{1,n} \\ ... & ... & ... \\ x_{n,1} & ... & x_{n,n} \end{bmatrix}}$$

We can see that we're ending up with the values of A, however they're transposed because of our trace summation identity.

So the answer is

$$A^T$$

6.4 Part d

$$a^T X b = \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ji} * a_i * b_j$$

$$\frac{\partial \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ji} * a_i * b_j}{\partial \begin{bmatrix} x_{1,1} & ... & x_{1,n} \\ ... & ... & ... \\ x_{m,1} & ... & x_{m,n} \end{bmatrix}}$$

Since we know that the output matrix has to be m x n and we derive out component by component we end up with a tensor or outer product of a and b. However since they are of dimensions m x 1 and n x 1 respectively, we need to transpose one in order to make the format work. Therefore we end up with

$$ab^T$$

6.5 Part e

To prove $||\mathrm{x}\ ||_2 \leq ||\mathrm{x}||_1$ Given x's greater than or equal to 1, less than or equal to 1, or equal to 0. The root of the sum of squares is equal to the sum of absolute values.

Given x's less than 1 and greater than -1 and not equal to 0. The root of the sum of squares is always less than the sum of absolute values because $\mathrm{x}^2 < \mathrm{x}$ given the aforementioned x requirements.

Therefore x is any and all real numbers, any combination will make it so that $||\mathrm{x}\ ||_2 \leq ||\mathrm{x}||_1$.

---

7 Problem 7

7.1 Part a

$\mathrm{R[w]} = \Lambda\ (\mathrm{Xw - y})^2$

7.2 Part b

In this part, X' == X^T, the ' is equal to the transpose.

$\mathrm{R[w]} = \Lambda\ (\mathrm{Xw - y})^2$

$= (\mathrm{Xw - y})^{\mathrm{T}}\ \Lambda\ (\mathrm{Xw - y})$

= w'X' Λ XW - y' Λ Xw - y' Λ y - w'X' Λ y

we can transpose the last term to get the middle term

= w'X' Λ XW - 2y' Λ Xw - y' Λ y

Now we can take the derivative w.r.t the weights of the above statement and set it to 0 to minimize the risk.

$0 = \frac{\partial(w'X'\Lambda XW - 2y'\Lambda Xw - y'\Lambda y)}{\partial w}$

$0 = \frac{\partial w'X'\Lambda Xw}{\partial w} - 2\frac{\partial y'\Lambda Xw}{\partial w} - \frac{\partial y'\Lambda y}{\partial w}$

remove the last term...

$0 = \frac{\partial w'X'\Lambda Xw}{\partial w} - 2\frac{\partial y'\Lambda Xw}{\partial w}$

now set β = X' Λ X and and α = y' Λ X

$0 = \frac{\partial w\beta w}{\partial w} - 2\frac{\partial \alpha w}{\partial w}$

from our identities and the previous problem we can derive these easily.

$0 = (\beta + \beta^{\mathrm{T}})\mathrm{w}$ - 2 α^{T}

now we can fill back in the β and α

0 = 2X' Λ Xw - 2X' Λ y

X' Λ y = X' Λ Xw

$\mathrm{w} = (\mathrm{X'}\ \Lambda\ \mathrm{X})^{-1}\ \mathrm{X'}\ \Lambda\ \mathrm{y}$

7.3 Part c

R[w] = Λ (Xw - y)^2 + γ w'w

Now we just need to add in another term to the end of our derivation.

0 = 2X' Λ Xw - 2X' Λ y + $\frac{\partial \gamma w' w}{\partial w}$

0 = 2X' Λ Xw - 2X' Λ y + 2 γ w

remove all 2's

X' Λ y = X' Λ Xw + γ w

we can bring in the identity matrix because $AI = A$

X' Λ y = (X' Λ X + γ I)w

w = (X' Λ X + γ I)$^{-1}$ X' Λ y

7.4 part d

8 Problem 8

8.1 Part a

8.2 Part b

If there is no loss when choosing doubt, the risk minimization that takes place will mean that the lowest risk choice is always doubt - making our algorithm useless. If the doubt cost is greater than the mis-classification cost then our algorithm will always take a guess at classifying the example and will never go with doubt. This is a way of forcing our algorithm to take a guess and what class something should belong to.