Advanced Machine Learning
Problem Set #0

Si Kai Lee sl3950@columbia.edu

September 12, 2016

Problem 1

I am hoping to do a PhD in ML working on linking traditional ML to Deep Learning. Doing CS/ML theory would give me a much stronger mathematical background which would definitely be useful for reading papers and the potential PhD.

Problem 2

a

False

b

False

Problem 3

a

$$A = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.125 & 0.125 & 0 \\ 0 & 0 & 0 & 0.125 & 0.125 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 \end{bmatrix}$$ R = 5 as rows 4 and 5 of A are the same.

b

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 2 \end{bmatrix}$$ is an eigenvector of A as its dot product with A yields $$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.25 \\ 0.25 \\ 0.5 \end{bmatrix}$$ which is 0.25 * B.

c

False. $$(A - 0.5I)X = \begin{bmatrix} -1.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.375 & 0.125 & 0 \\ 0 & 0 & 0 & 0.125 & -0.375 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.25 \end{bmatrix} X = 0.$$

Solving the set of linear equations above, we find that the 2nd and 3rd elements of the eigenvector are free variables while the remaining elements equate to 0.

Thus, the eigenvectors are of the form $$\begin{bmatrix} 0 \\ c \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$ and $$\begin{bmatrix} 0 \\ 0 \\ d \\ 0 \\ 0 \\ 0 \end{bmatrix}$$ as the 2nd and 3rd elements are free variables.

d

From Σ of A, we know two eigenvectors with eigenvalue 0.25 exist, so the dimension of $V = 2$.

$$(A - 0.25I)X = \begin{bmatrix} -1.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.125 & 0.125 & 0 \\ 0 & 0 & 0 & 0.125 & -0.125 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} X = 0.$$

Through solving the set of linear equations above, we find the 4th and 5th elements of eigenvector X_1 to be equal and the 6th element of X_2 a free variable which confirms that $V = 2$.

e

Since A is expressed in the form $Q\Sigma Q^{-1}$ where the eigenvalues of A are the diagonals in Σ. Taking Σ^3, we find that the largest eigenvalue of A^3 is $0.5^3 = 0.125$.

Problem 4

a

Assuming a negative binomial distribution over coin tosses, we have $\mathbb{E}(x) = \frac{1-p}{p}$ where p is the probability of seeing heads. Since p $= \frac{1}{5}$ so $\mathbb{E}$(# tails seen before first heads) $= \frac{4}{5}/\frac{1}{5} = 4$. Since we have to include the actual toss that leads to heads, $\mathbb{E}$(# tosses to see heads) $= 4 + 1 = 5$.

b [1]

Assuming a Bernoulli distribution over n words, we have $\mathbb{E}(x) = np$ where n is the # of phrases of length 5 and p the probability of seeing 'a rose is a rose'. We know that within n words, there are $n - 5$ different sequences of length 5 and the probability of seeing the phrase 'a rose is a rose' is $\frac{1}{4}^5$. Hence $\mathbb{E}$(# 'a rose is a rose' in n words) $= (n-5) * \frac{1}{4}^5$

c

Assuming a uniform discrete distribution over T, $\mathbb{E}(x^2) = \sum_{x \in \Omega} x^2 \mathrm{P}(X = x)$. We also assume that there are n different permutations of t, then $\mathbb{E}(x^2) = \sum_{x=1}^{n} x^2 \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$.

d

To obtain the CDF of the given distribution, we integrate it from 0 to ∞. $\int_0^{1000000} \lambda e^{-\lambda x} = [-e^{-\lambda x}]_0^{1000000} = 1 - e^{-1000000\lambda} = 0.5$. Hence, by moving 1 to the RHS, multiplying by -1 and taking logs, we have $-1000000\lambda = \ln(0.5)$ which gives us $\lambda = -\frac{\ln(0.5)}{1000000}$.

e

For p to be a valid PDF, $\int_0^{0.5} \int_0^1 c dx_1 dx_2 = 1$. Integrating the LHS, we have $0.5c = 1$, hence $c = 2$.

f

$P(X_2 \geq X_1) = \int_0^{0.5} \int_{x_1}^1 2 dx_2 dx_1 = 2 \int_0^{0.5} (1 - x_1) dx_1 = 2[x_1 - 0.5{x_1}^2]_0^{0.5} = 0.75$.

g

No. $P(X_1 > 2X_2) = P(X_1 \geq 2X_2) = \int_0^{0.25} \int_0^{2x_2} 2 dx_1 dx_2 = 2 \int_0^{0.25} 2x_2 dx_2 = 2[{x_2}^2]_0^{0.25} = 0.125$. Hence, $\mathbb{E}(Y) = 0.125 * 1 + 0.875 * -1 = -0.75$.

h

Yes. $P(X_2 > 0.5) = P(X_2 \geq 0.5) = \int_{0.5}^1 \int_0^{0.5} 2 dx_1 dx_2 = \int_{0.5}^1 dx_2 = 0.5$. Since X_1 and Z are independent, $\mathbb{E}(X_1 Z) = \mathbb{E}(X_1)\mathbb{E}(Z) = 0.25 * 0 = 0$.

[1] Discussed with Maja Ruldoph mrr2163

Problem 5

a

We know that $q \in \text{conv}(S)$ is equivalent to q being a combination of $x_i \in S$ as defined by $\text{conv}(S)$. In addition, subtracting q from r and x is a centring operation that centres r and x around q.

Since the convex hull of S $\text{conv}(S)$ is defined as the envelope of the convex set of S, we have for every pair of points within S, every point on the straight line segment that joins the pair of points is also within S and that is true for all d-dimensions covered by S. Following the above, there must be some $x \in S$ that is pointing at least 90 degrees away from r. Hence for any $r \in \mathbb{R}^d$, there exist $x \in S$ such that $\langle r - q, x - q \rangle \leq 0$.

b [2]

We know that $\Delta^2 = \max_{x,y \in S} ||x - y||_2^2$. To obtain the above, y must be the combination of vectors that point the direction farthest away from x, i.e. $y \approx -x$. Also, we can be sure that $||z - y||_2^2 < \Delta^2$ as z and y are within the convex hull itself and Δ^2 represents the maximum distance between the vertices of the the convex hull.

We can rewrite $||z - y||_2^2$ as $||y - z||_2^2$. By substituting the definition of y and expanding the 2-norm, we have $\langle \frac{1}{T} \sum_{i=1}^T x_i - z, \frac{1}{T} \sum_{i=1}^T x_i - z \rangle = \frac{1}{T^2} \langle \sum_{i=1}^T x_i - Tz, \sum_{i=1}^T x_i - Tz \rangle$. We now split $\sum_{i=1}^T x_i$ into $\sum_{i=1}^{T-1} x_i + x_T$ and get $\frac{1}{T^2} \langle (\sum_{i=1}^{T-1} x_i + x_T - Tz), (\sum_{i=1}^{T-1} x_i + x_T - Tz) \rangle$. Further expanding the above, we get

$\frac{1}{T^2}(\langle x_T - z, x_T - z \rangle + 2\langle \sum_{i=1}^{T-1} -(T-1)z, x_T - z \rangle + \langle \sum_{i=1}^{T-1} -(T-1)z, \sum_{i=1}^{T-1} -(T-1)z \rangle)$

We know that $\langle x_T - z, x_T - z \rangle \leq \Delta^2$ and $\langle \sum_{i=1}^{T-1} -(T-1)z, x_T - z \rangle \leq 0$ so the above equation is less than $\frac{1}{T^2}(\Delta^2 + \langle \sum_{i=1}^{T-1} -(T-1)z, \sum_{i=1}^{T-1} -(T-1)z \rangle)$. By splitting the $\sum_{i=1}^T x_i - Tz$ T times, we get $||y - z||_2^2 \leq \frac{1}{T^2} T \Delta^2 = \Delta^2 / T$.

[2]Discussed with Erik Waingarten eaw2197