CS6140: Machine Learning

Homework Assignment # 1

Assigned: 01/21/2020 *Due: 02/04/2020, 11:59pm, through Blackboard*

Submitted by: Piyush Goel
goel.pi@husky.neu.edu

Problem 1. (5 points) Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A \subseteq \Omega$ and $B \subseteq \Omega$ any two subsets of Ω. Prove the following expression or provide a counterexample if it does not hold

$$P(A) = P(A|B) + P(A|B^c),$$

where A^c is the complement of A.

Solution
By the law of total partition (sum rule), we know that

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Hence the equation given is not valid. Here is a counter example as well.
Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event A be that a randomly chosen number out of Ω is divisible by 3. The event B be that a randomly chosen number out of Ω is divisible by 2.
Therefore, $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = 1/3$.
Similarly, $P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{1/6}{1/2} = 1/3$.
So, $P(A|B) + P(A|B^c) = 2/3 \neq P(A) = 1/3$.

Problem 2. (15 points) Let X be a random variable on $\mathcal{X} = \{a, b, c\}$ with the probability mass function $p(x)$. Let $p(a) = 0.1$, $p(b) = 0.2$, and $p(c) = 0.7$ and some function $f(x)$ be

$$f(x) = \begin{cases} 10 & x = a \\ 5 & x = b \\ \frac{10}{7} & x = c \end{cases}$$

a) (5 points) What is $\mathbb{E}[f(X)]$?

b) (5 points) What is $\mathbb{E}[1/p(X)]$?

c) (5 points) For an arbitrary finite set $\mathcal{X}$ with n elements and arbitrary $p(x)$ on $\mathcal{X}$, what is $\mathbb{E}[1/p(X)]$?

Solution

a) $\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} p_X(x) f(x) = p(a)f(a) + p(b)f(b) + p(c)f(c) = 3$

b) $\mathbb{E}[1/p(X)] = \sum_{x \in \mathcal{X}} p_X(x)/p_X(x) = \sum_{x \in \mathcal{X}} 1 = 3$

c) $\mathbb{E}[1/p(X)] = \sum_{x\in\mathcal{X}} p_X(x)/p_X(x) = \sum_{x\in\mathcal{X}} 1 = |\mathcal{X}|$ (i.e. the cardinalty of the set $\mathcal{X}$)

Problem 3. (10 points) Let X and Y be random variables. Prove or disprove the following formula

$$V[X+Y] = V[X] + V[Y] + 2\text{Cov}[X,Y],$$

where $V[X]$ is the variance of X and $\text{Cov}[X,Y]$ is the covariance between X and Y.

Solution

$$V[X] = \mathbb{E}[(x - \mathbb{E}[X])^2] = \mathbb{E}[x^2 - 2x\mathbb{E}[X] + \mathbb{E}[X]^2] = \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Similarly,

$$V[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$$

We also know,

$$\text{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Therefore,

$$V[X] + V[Y] + 2\text{Cov}[X,Y] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y]$$

(Using $c\mathbb{E}[X] = \mathbb{E}[cX]$ and $\mathbb{E}[X] + \mathbb{E}[Y] = \mathbb{E}[X+Y]$)

$$= \mathbb{E}[X^2 + Y^2 + 2XY] + \mathbb{E}[X]^2 + \mathbb{E}[Y]^2 + 2\mathbb{E}[X]\mathbb{E}[Y]$$

$$= \mathbb{E}[(X+Y)^2] + (\mathbb{E}[X] + \mathbb{E}[Y])^2 = V[X+Y]$$

Hence proved $V[X+Y] = V[X] + V[Y] + 2\text{Cov}[X,Y]$.

Problem 4. (20 points) Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with an unknown mean λ. Based on previous experience in similar industrial plants, suppose that our initial feelings about the possible value of λ can be expressed by an exponential distribution with parameter $\theta = \frac{1}{2}$ is, the prior density is

$$p(\lambda) = \theta e^{-\theta\lambda}$$

where $\lambda \in (0, \infty)$. If there are 99 accidents over the next 11 days, determine

a) (5 points) the maximum likelihood estimate of λ

b) (5 points) the maximum a posteriori estimate of λ

c) (10 points) the Bayes estimate of λ.

Solution

a)

$$\lambda_{ML} = \underset{\lambda}{\text{argmax}}\{p(D|\lambda)\} = \underset{\lambda}{\text{argmax}}\{\log p(D|\lambda)\} = \underset{\lambda}{\text{argmax}}\{\log \prod_{x\in D} p(x|\lambda)\} = \underset{\lambda}{\text{argmax}}\{\sum_{x\in D} \log p(x|\lambda)\}$$

$$= \underset{\lambda}{\text{argmax}} \left\{ \sum_{x \in D} \log \left(\frac{\lambda^x e^{-\lambda}}{x!} \right) \right\} = \underset{\lambda}{\text{argmax}} \left\{ \sum_{x \in D} x \log \lambda - \lambda - \log x! \right\}$$

Now,

$$\frac{\partial}{\partial \lambda} \left(\sum_{x \in D} x \log \lambda - \lambda - \log x! \right) = \sum_{x \in D} \frac{x}{\lambda} - 1$$

Now we can set the derivative equal to 0 for the maximization.

$$\sum_{x \in D} \frac{x}{\lambda_{ML}} - 1 = 0$$

Hence,

$$\lambda_{ML} = \frac{\sum_{x \in D} x}{\sum_{x \in D} 1} = \frac{99}{11} = 9$$

b)

$$\lambda_{MAP} = \underset{\lambda}{\text{argmax}} \{ p(\lambda) p(D|\lambda) \} = \underset{\lambda}{\text{argmax}} \{ \log p(\lambda) p(D|\lambda) \} = \underset{\lambda}{\text{argmax}} \{ \log p(\lambda) + \log p(D|\lambda) \}$$

Considering just $\log p(\lambda)$

$$\log p(\lambda) = \log \theta e^{-\theta \lambda} = \log \theta - \theta \lambda$$

We already have $\log p(D|\lambda)$ from the previous part of the question. Therefore,

$$\log p(\lambda) + \log p(D|\lambda) = \log \theta - \theta \lambda + \sum_{x \in D} x \log \lambda - \lambda - \log x!$$

Taking its partial derivative, with respect to λ and setting it equal to zero we get,

$$-\theta + \frac{1}{\lambda_{MAP}} \sum_{x \in D} x - \sum_{x \in D} 1 = 0$$

Hence,

$$\lambda_{MAP} = \frac{\sum_{x \in D} x}{\theta + \sum_{x \in D} 1} = \frac{99}{11.5} = 8.60869$$

c)

$$\lambda_B = \int_{\lambda \in (0,\infty)} \lambda p(\lambda|D) d\lambda$$

$$p(\lambda|D) = \frac{p(D|\lambda) p(\lambda)}{p(D)}$$

$$p(D|\lambda) = \prod_{x \in D} p(x|\lambda) = \prod_{x \in D} \frac{\lambda^x e^{-\lambda}}{x!} = \frac{\lambda^{\sum_{x \in D} x} e^{-\lambda n}}{\prod_{x \in D} x!}$$

$$p(D) = \int_{\lambda \in (0,\infty)} p(D|\lambda) p(\lambda) = \int_{\lambda \in (0,\infty)} \frac{\lambda^{\sum_{x \in D} x} e^{-\lambda n}}{\prod_{x \in D} x!} \theta e^{-\theta \lambda} d\lambda = \frac{\theta}{\prod_{x \in D} x!} \int_{\lambda \in (0,\infty)} \lambda^{\sum_{x \in D} x} e^{-(n+\theta)\lambda} d\lambda$$

$$= \frac{\theta}{\prod_{x \in D} x!} \frac{\Gamma(1 + \sum_{x \in D} x)}{(n+\theta)^{1+\sum_{x \in D}}}$$

Therefore

$$\lambda_B = \int_{\lambda \in (0,\infty)} \lambda \frac{\lambda^{\sum_{x \in D} x} e^{-\lambda n}}{\prod_{x \in D} x!} \theta e^{-\theta \lambda} \frac{\prod_{x \in D} x!}{\theta} \frac{(n+\theta)^{1+\sum_{x \in D}}}{\Gamma(1 + \sum_{x \in D} x)}$$

$$= \frac{(n+\theta)^{1+\sum_{x \in D} x}}{\Gamma(1 + \sum_{x \in D} x)} \frac{\Gamma(2 + \sum_{x \in D} x)}{(n+\theta)^{2+\sum_{x \in D} x}} = \frac{1 + \sum_{x \in D} x}{n + \theta} = \frac{1 + 99}{11 + 0.5} = 8.69565$$

Problem 5. (10 points) Let $\mathcal{D} = \{x_i\}_{i=1}^n$ be an i.i.d. sample from

$$p(x) = \begin{cases} e^{-(x-\theta_0)} & x \geq \theta_0 \\ 0 & \text{otherwise} \end{cases}$$

Determine θ_{ML} – the maximum likelihood estimate of θ_0.

Solution
Using θ instead of θ_0 just for convenience of notation.

$$\theta_{ML} = \underset{\theta}{\text{argmax}}\{p(D|\theta)\} = \underset{\theta}{\text{argmax}}\{\log p(D|\theta)\} = \underset{\theta}{\text{argmax}}\{\log \prod_{x \in D} p(x|\theta)\} = \underset{\theta}{\text{argmax}}\{\sum_{x \in D} \log p(x|\theta)\}$$

Ignoring the constraint on x for now we get,

$$\sum_{x \in D} \log p(x|\theta) = \sum_{x \in D} \log e^{-(x-\theta)} = \sum_{x \in D} \theta - x$$

Now, we have a constrained maximization here. So, the new function to be maximized is

$$L(\theta, \mu_1, \mu_2, \cdots \mu_n) = \left(\sum_{x \in D} \theta - x \right) + \sum_{k=1}^{n} \mu_k (x_k - \theta)$$

Such that $x_k \geq \theta$, $\mu_k \geq 0$ and $\mu_k(x_k - \theta) = 0 \ \forall k \in (1, n)$. Now, taking the partial derivative of L with respect to θ and setting it equal to 0 we get,

$$\sum_{x \in D} 1 - \sum_{k=1}^{n} \mu_k = 0 \implies \sum_{k=1}^{n} \mu_k = n$$

So, now $\theta \leq x_k \forall k \in (1, n)$, or $\theta \leq x_{min}$, where x_{min} is the smallest of all the $x'es$. Hence, $\theta_{ML} = x_{min}$.
Another Reasoning:
It can also be observed that $p(x_k) \forall k \in [1, n]$ can have only amongst two possible values, either 0 or $e^{-x_k+\theta}$ (> 0) depending on whether $x \geq \theta$. Now, to maximize the likelyhood of the data, we can maximize the likelyhood of all datapoints, since the data is i.i.d. and the likelyhood of any datapoint x_k would be max if $x_k \geq \theta$. Now, for the total likelyhood to be max $x_1, x_2, \cdots x_n \geq \theta$ should be true, or simply $x_{min} \geq \theta$, where x_{min} is the smallest of all the $x'es$. Therefore we arrive at the same result using this argument as well, that is $\theta_{ML} = x_{min}$.

Problem 6. (25 points) Let $\mathcal{D} = \{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}$, be a data set drawn independently from a Gumbel distribution

$$p(x) = \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} e^{-e^{-\frac{x-\alpha}{\beta}}},$$

where $\alpha \in \mathbb{R}$ is the location parameter and $\beta > 0$ is the scale parameter.

a) (10 points) Derive an algorithm for estimating α and β.

b) (10 points) Implement the algorithm derived above and evaluate it on data sets of different sizes. First, find or develop a random number generator that creates a data set with $n \in \{100, 1000, 10000\}$ values using some fixed α and β. Then make at least 10 data sets for each n and estimate the parameters. For each n, report the mean and standard deviation on the estimated α and β. If $n = 10000$ is too large for your computing resources, skip it.

c) (5 points) The problem above will require you to implement an iterative estimation procedure. You will need to decide on how to initialize the parameters, how to terminate the estimation process and what the maximum number of iterations should be. Usually, some experimentation will be necessary *before* you run the experiments in part (b) above. Summarize what you did in a short paragraph, no more than two paragraphs.

NB: There are several versions and naming conventions for the Gumbel distribution in the literature.
Solution

a) Using the maximum likelyhood estimates.

$$\alpha_{ML} = \underset{\alpha}{\operatorname{argmax}} \left\{ \sum_{x \in D} \log p(x|\alpha, \beta) \right\}, \beta_{ML} = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{x \in D} \log p(x|\alpha, \beta) \right\}$$

Let $f = \sum_{x \in D} \log p(x|\alpha, \beta) = \sum_{x \in D} -\log \beta - \frac{x-\alpha}{\beta} - e^{-\frac{x-\alpha}{\beta}}$.
It is not easy to find the optimum α and β due to the complicated derivatives. So, we will use Newton-Raphson optimization to numerically approximate the optimums.
Let the vector θ^t be $\begin{bmatrix} \alpha^t \\ \beta^t \end{bmatrix}$, i.e. the value of α and β at the t^{th} iteration of Newton-Raphson.
Let ∇ be the vector of the first derivatives.

$$\nabla = \begin{bmatrix} \frac{\partial f}{\partial \alpha} \\ \frac{\partial f}{\partial \beta} \end{bmatrix} = \begin{bmatrix} \sum_{x \in D} \frac{1}{\beta} - \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} \\ \sum_{x \in D} -\frac{1}{\beta} + \frac{x-\alpha}{\beta^2} - \frac{x-\alpha}{\beta^2} e^{-\frac{x-\alpha}{\beta}} \end{bmatrix}$$

Let H be the Hessian Matrix of double derivatives.

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial^2 \alpha} & \frac{\partial^2 f}{\partial \alpha \partial \beta} \\ \frac{\partial^2 f}{\partial \beta \partial \alpha} & \frac{\partial^2 f}{\partial^2 \beta} \end{bmatrix} = \begin{bmatrix} \sum_{x \in D} -\frac{1}{\beta^2} e^{-\frac{x-\alpha}{\beta}} & \sum_{x \in D} -\frac{1}{\beta^2} + \frac{1}{\beta^2} e^{-\frac{x-\alpha}{\beta}} - \frac{x-\alpha}{\beta^3} e^{-\frac{x-\alpha}{\beta}} \\ \sum_{x \in D} -\frac{1}{\beta^2} + \frac{1}{\beta^2} e^{-\frac{x-\alpha}{\beta}} - \frac{x-\alpha}{\beta^3} e^{-\frac{x-\alpha}{\beta}} & \sum_{x \in D} \frac{1}{\beta^2} - \frac{2(x-\alpha)}{\beta^3} + \frac{2(x-\alpha)}{\beta^3} e^{-\frac{x-\alpha}{\beta}} + (\frac{x-\alpha}{\beta^2})^2 e^{-\frac{x-\alpha}{\beta}} \end{bmatrix}$$

Now we can simply start from any θ^0 and update it using the rule $\theta^{t+1} = \theta^t - H^{-1}\nabla$, until it converges.

b) The values of [mean(α), std-dev(α), mean(β), std-dev(β)], calculated over 10 different random datasets (with the true values, $\alpha = 3, \beta = 5$) respectively are
For $n = 100$, $[3.048, 0.372, 4.983, 0.402]$
For $n = 1000$, $[3.035, 0.221, 4.950, 0.126]$
For $n = 10000$, $[3.006, 0.039, 5.018, 0.042]$

c) The problem faced while initializing the parameters for the algorithm is that if we start off too far from the actual values of the parameters, the algorithm did not converge, and the values of the parameters seem to blow up to ∞.
The estimation process was terminated when the parameters estimated at step t are very close to the parameters estimated at step $t+1$, i.e. if their difference is less than a certain value, say γ. The value of γ didn't have a lot of difference on the finally calculated values, given that it is small enough (around 0.0001).
The number of steps is limited to be 100, but that doesn't really affect the final value of the parameters, since the algorithm converges with much less iterations, given the parameters are initialized properly.

Problem 7. (30 points) Let $\mathcal{D} = \{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}$, be a data set drawn independently from the mixture of two distributions

$$p(x) = w_1 p_1(x) + w_2 p_2(x),$$

where

$$p_1(x) = \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} e^{-e^{-\frac{x-\alpha}{\beta}}},$$

is a Gumbel distribution with parameters α and β,

$$p_2(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

is a Gaussian distribution with parameters μ and σ, and (w_1, w_2) are positive constants such that $w_1 + w_2 = 1$.

a) (15 points) Derive an EM algorithm for estimating w_1, w_2, α, β, μ and σ.

b) (15 points) Implement the algorithm derived above and evaluate it on data sets of different sizes. Use similar experimentation as in Problem 3.

You can recycle derivations and code from Problem 6. You can also use any result and derivation from the lecture notes posted on the class web site.

Solution
Let $\theta = \begin{bmatrix} w_1 & w_2 & \alpha & \beta & \mu & \sigma \end{bmatrix}$, i.e. a vector of parameters.

From the class notes, we already know that

$$E[\log p(D, Y|\theta)|D, \theta^{(t)}] = \sum_{i=1}^{n} \sum_{j=1}^{m} \log(w_j p(x_i|\theta_j)) p(y_i = j|x_i, \theta^{(t)})$$

Let $f = E[\log p(D, Y|\theta)|D, \theta^{(t)}]$.

$$f = \sum_{i=1}^{n} \log(w_1 p(x_i|\theta_1)) p(y_i = 1|x_i, \theta^{(t)}) + \log(w_2 p(x_i|\theta_2)) p(y_i = 2|x_i, \theta^{(t)})$$

Taking derivative with respect to w_1 and w_2 (while considering the constraints using Lagrange multipliers) and setting it to zero, we get (as done in class)

$$w_1 = \frac{1}{n} \sum_{i=1}^{n} p(y_i = 1|x_i, \theta^{(t)}), w_2 = \frac{1}{n} \sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)})$$

Now, taking the partial derivative of f with respect to α we get

$$\frac{\partial f}{\partial \alpha} = \sum_{i=1}^{n} p(y_i = 1|x_i, \theta^{(t)}) \frac{\partial}{\partial \alpha} \log w_1 p(x_i|\theta_1) = \sum_{i=1}^{n} p(y_i = 1|x_i, \theta^{(t)}) \frac{\partial}{\partial \alpha} \log w_1 + \log p(x_i|\theta_1)$$

Using the derivatives from the previous problem

$$\frac{\partial f}{\partial \alpha} = \sum_{i=1}^{n} p(y_i = 1|x_i, \theta^{(t)}) \frac{\partial}{\partial \alpha} \log p(x_i|\theta_1) = \sum_{i=1}^{n} p(y_i = 1|x_i, \theta^{(t)}) \left(\frac{1}{\beta} - \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} \right)$$

Similarly,

$$\frac{\partial f}{\partial \beta} = \sum_{i=1}^{n} p(y_i = 1|x_i, \theta^{(t)}) \left(-\frac{1}{\beta} + \frac{x-\alpha}{\beta^2} - \frac{x-\alpha}{\beta^2} e^{-\frac{x-\alpha}{\beta}} \right)$$

The hessian matrix can similarly be calculated as

$$\begin{bmatrix} \sum_{x \in D} p(y_i = 1|x_i, \theta^{(t)}) \left(-\frac{1}{\beta^2} e^{-\frac{x-\alpha}{\beta}} \right) & \sum_{x \in D} p(y_i = 1|x_i, \theta^{(t)}) \left(-\frac{1}{\beta^2} + \frac{1}{\beta^2} e^{-\frac{x-\alpha}{\beta}} - \frac{x-\alpha}{\beta^3} e^{-\frac{x-\alpha}{\beta}} \right) \\ \sum p(y_i = 1|x_i, \theta^{(t)}) \left(-\frac{1}{\beta^2} (1 - e^{-\frac{x-\alpha}{\beta}}) - \frac{x-\alpha}{\beta^3} e^{-\frac{x-\alpha}{\beta}} \right) & \sum p(y_i = 1|x_i, \theta^{(t)}) \left(\frac{1}{\beta^2} - \frac{2(x-\alpha)}{\beta^3} (1 - e^{-\frac{x-\alpha}{\beta}}) + (\frac{x-\alpha}{\beta^2})^2 e^{-\frac{x-\alpha}{\beta}} \right) \end{bmatrix}$$

Removed the $x \in D$ from the last two summations, just to fit the matrix in the page.

Similarly, for μ and σ

$$\frac{\partial f}{\partial \mu} = \sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)}) \frac{1}{p(x_i|\theta_2)} \frac{\partial}{\partial \mu} p(x_i|\theta_2) = \sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)}) \frac{x-\mu}{\sigma^2}$$

$$\frac{\partial f}{\partial \sigma} = \sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)}) \frac{1}{p(x_i|\theta_2)} \frac{\partial}{\partial \sigma} p(x_i|\theta_2) = \sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)}) \frac{1}{\sigma^3} \left((x-\mu)^2 - \sigma^2 \right)$$

Setting these derivatives equal to 0 gives us

$$\mu^* = \frac{\sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)}) x_i}{\sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)})}$$

$$\sigma^{*2} = \frac{\sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)}) (x-\mu^*)^2}{\sum_{i=1}^{n} p(y_i = 2|x_i, \theta^{(t)})}$$

Now, the EM algorithm is as follows:

a) Initalialize $\theta^{(0)}$ (i.e. $\alpha^{(0)}$, $\beta^{(0)}$, $\mu^{(0)}$, $\sigma^{2(0)}$, $w_1^{(0)}$ and $w_2^{(0)}$).

b) Set t = 0.

c) Repeat until convergence:

   (a) $p(y_i = 1|x_i, \theta^{(t)}) = \frac{w_1 p(x_i|\alpha^{(t)}, \beta^{(t)})}{w_1 p(x_i|\alpha^{(t)}, \beta^{(t)}) + w_2 p(x_i|\mu^{(t)}, \sigma^{2(t)})}$, $p(y_i = 2|x_i, \theta^{(t)}) = \frac{w_2 p(x_i|\mu^{(t)}, \sigma^{2(t)})}{w_1 p(x_i|\alpha^{(t)}, \beta^{(t)}) + w_2 p(x_i|\mu^{(t)}, \sigma^{2(t)})}$

   (b) Update w_1 and w_2 using the equations derived above.

   (c) Update α and β using newton raphson as discussed above, and update μ and σ using their respective optimums as derived above.

   (d) $t = t + 1$.

d) Report $\theta^{(t)}$.

The following means and standard deviations were observed after doing the required experimentation, written in the following order: $[mean(\alpha), stddev(\alpha), mean(\beta), stddev(\beta), mean(\mu), stddev(\mu), mean(\sigma), stddev(\sigma), mean(w_1), stddev(w_1), mean(w_2), stddev(w_2)]$
For $n = 100$ $[3.653, 2.976, 4.313, 0.687, 1.035, 0.540, 1.633, 0.583, 0.506, 0.235, 0.493, 0.235]$
For $n = 1000$ $[3.258, 0.825, 4.820, 0.177, 0.937, 0.092, 1.939, 0.102, 0.453, 0.076, 0.546, 0.076]$
For $n = 10000$ $[3.170, 0.199, 5.087, 0.105, 0.998, 0.040, 2.026, 0.054, 0.431, 0.018, 0.568, 0.018]$

Problem 8. (20 points) Understanding the curse of dimensionality. Consider the following experiment: generate n data points with dimensionality k. Let each data point be generated using a uniform random number generator with values between 0 and 1. Now, for a given k, calculate

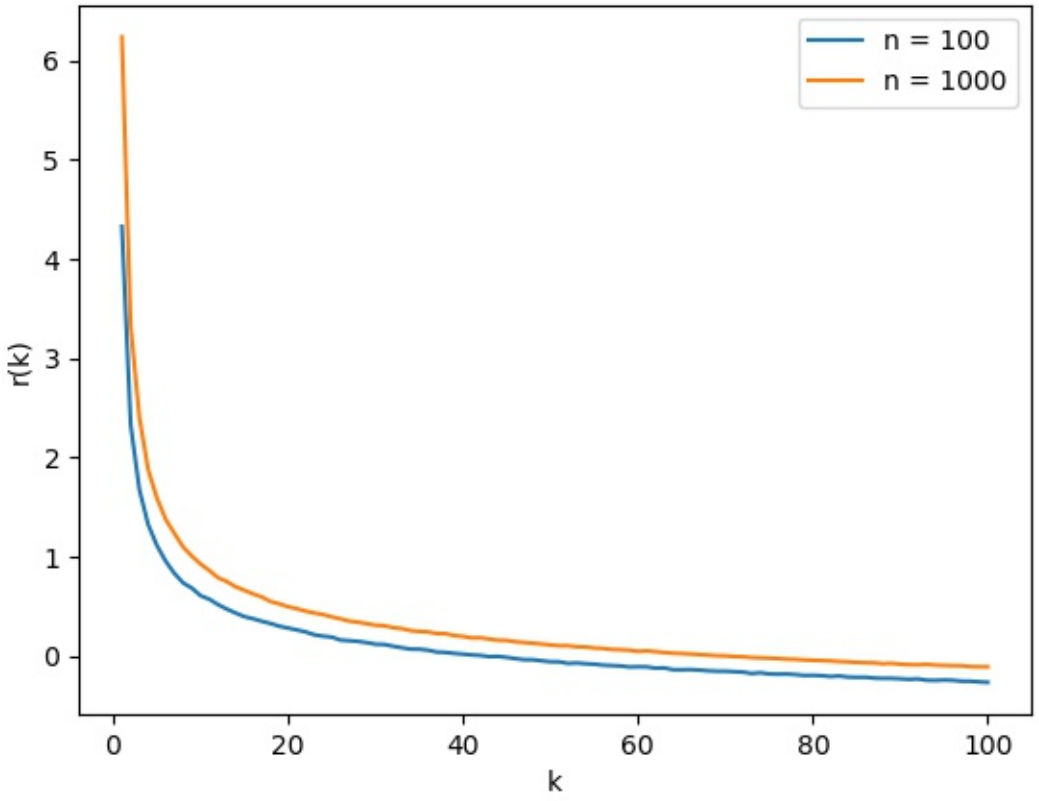
$$r(k) = \log_{10} \frac{d_{\max}(k) - d_{\min}(k)}{d_{\min}(k)}$$

where $d_{\max}(k)$ is the maximum distance between any pair of points and $d_{\min}(k)$ is minimum distance between any pair of points (you cannot use identical points to obtain the minimum distance of 0). Let k take each value from $\{1, 2, \ldots, 99, 100\}$. Repeat each experiment multiple times to get stable values by averaging the quantities over multiple runs for each k.

a) (15 points) Plot $r(k)$ as a function of k for two different values of n; $n \in \{100, 1000\}$. Label and scale each axis properly to be able to make comparisons over different n's. Embed your final picture(s) in the file you are submitting for this assignment.

b) (5 points) Discuss your observations and also compare the results to your expectations before you carried out the experiment.

Solution



Expectation before carrying out the experiment was that the value of $r(k)$ would stay close to constant as k increases, thinking that the distance between the closest pairs of points would increase at the same rate

as the distance between the farthest pair of points grow and hence cancel out each other's effect since we're finally taking their ratios. It was also expected that increasing the number of points would significantly increase the value of the function for the same k because the distance between the closest pair of points should be decreasing with increasing n but the distance between the farthest pair of points should stay approximately the same.
The **observation** was quite different from what was expected, the value of $r(k)$ decreased with the increase in k. And the increase in $r(k)$ wasn't much after increasing n 10 folds, but this may have been due to the log.