CS6140: Machine Learning

Homework Assignment # 1

Assigned: 09/18/2022 *Due: 10/03/2022, 11:59pm, through Canvas*

Eight problems, 170 points in total. Good luck!
Prof. Predrag Radivojac, Northeastern University

Problem 1. (20 points) Two players perform a series of coin tosses. Player one wins a toss if the coin turns heads and player two wins a toss if it turns tails. The game is played until one player wins n times. However, the game is interrupted when player one had l wins and player two had m wins, where $0 \leq l, m < n$.

a) (5 points) Assuming $n = 8$, $l = 4$, and $m = 6$, what is the probability that player one would win the game if the game was to be continued later.

b) (10 points) Derive the general expression or write an algorithm that computes the probability that player one will win the game if the game is to be continued later. Your expression should be a function of n, m, and l. If you are providing an algorithm, implement it and submit your code along with your pseudo-code that should be in your report. You may *not* simulate the game as your solution, though you can use simulation to verify your solution. Hint: negative binomial distribution may be useful.

c) (5 points) When l and m are kept constant, describe the influence of n to the final probability. How does that compare to what you thought about the problem before you solved it? Was your intuition right?

Answer.

a) When $n = 8, l = 4$ and $m = 6$, after the game is interrupted, player one has won 4 tosses, and player two has won 6 tosses. Once the game is continued, player one needs to win 4 more tosses before player two win 2 tosses in order to win.

Hence, the two player needs to play at least another 2 tosses and at most another 5 tosses in order for one player to win.

The possibility of playing i tosses such that we get exactly r heads can be calculated using the negative binomial distribution: $P(x) = \binom{i-1}{r-1} \cdot 0.5^r \cdot 0.5^{i-r}$

Using the formula above, we can calculate the possibility of player one winning by taking the sum of the chance of player one winning if they have to play exactly $2, 3, ..., 5$ games.

The possibility of player one winning is: $\sum_{i=0}^{5} \binom{i-1}{4-1} \cdot 0.5^4 \cdot 0.5^{i-4} = 0.1875 = 18.75\%$

b) Using the same approach as part a), we can calculate the possibility of player one win in the general case:

The number of tosses player one need to win the game is $r = n - l$. The number of tosses player two need to win is $n - m$. Hence, the two player need to play at most $n - m + n - l - 1 = 2n - m - l - 1$ tosses in order for one of them to win.

The possibility of player one winning is: $\sum_{i=0}^{2n-m-l-1} \binom{i-1}{n-l-1} \cdot 0.5^{n-l} \cdot 0.5^{i-(n-l)}$

c) If we keep m and l constant, as n increases to infinity, the possibility of player one win the game will be calculated as:

$$\begin{aligned} P(X) &= \lim_{n\to\infty} \sum_{i=0}^{2n-m-l-1} \binom{i-1}{n-l-1} \cdot 0.5^{n-l} \cdot 0.5^{i-(n-l)} \\ &\approx \lim_{n\to\infty} \sum_{i=0}^{2n} \binom{i-1}{n} \cdot 0.5^{n} \cdot 0.5^{i-n} \\ &= \lim_{n\to\infty} \sum_{i=0}^{2n} \binom{i-1}{n} \cdot 0.5^{i} && = 0.5 \end{aligned}$$

Hence, as n increases, we see that the chance for player one winning gets closer to 50%, which makes sense and adhere to my initial intuition.

Problem 2. (5 points) Let $(\Omega, \mathcal{A}, P)$ be a discrete probability space, where $\mathcal{A} = \mathcal{P}(\Omega)$, and let $A \subseteq \Omega$ and $B \subseteq \Omega$ be 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.

Answer.

Let consider the following situation: Let $P(B|\Omega) = P(B) = \frac{1}{4}$ and $P(A|\Omega) = P(A) = \frac{1}{8}$, A is completely within B (or $P(A \cap B^C) = 0$). Hence, $P(A|B) = \frac{1}{2}$.

We have:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{1/8}{1/4} = \frac{1}{2} \\ P(A|B^C) &= \frac{P(A \cap B^C)}{P(B^C)} = 0 \\ P(A|B) + P(A|B^C) &= \frac{1}{2} \neq P(A) \end{aligned}$$

Hence, the statement above is incorrect.

Problem 3. (25 points) Mary is going shopping for books to take on a trip. She will spend X hours in the bookstore, where X is a discrete random variable equally likely to take on values of 1, 2, 3, or 4. She will buy Y books, where Y is a discrete random variable that depends on the amount of time she shops as described by a conditional probability mass function

$$p(y|x) = \frac{1}{x}; \quad y = 1, 2, \ldots, x$$

a) (5 points) Find the joint probability mass function of X and Y.

b) (5 points) Find the marginal probability mass function for Y.

c) (5 points) Find the conditional probability mass function of X given that $Y = 2$

d) (5 points) Suppose you know that Mary bought at least two but not more than three books. Find the conditional mean and variance of X given this information.

e) (5 points) The cost of each book is a random variable (independent of all the other random variables mentioned) with mean 3. What is the total expected expenditure of Mary's visit to the bookstore?

Answer.

a) The formula for the joint probability mass function is: $P_{XY}(x,y) = P(X = x, Y = y)$

In this case, we have: $P(1,1) = \frac{1}{4}, P(2,1) = P(2,2) = \frac{1}{8}, P(3,1) = P(3,2) = P(3,3) = \frac{1}{12}, P(4,1) = P(4,2) = P(4,3) = P(4,4) = \frac{1}{16}$.

Which should satisfy: $\sum_{(x_i,y_i)\in\Omega_{XY}} P(x_i, y_i) = 1$.

b) The marginal probability mass function for Y, p_Y, is: $p_Y(y) = P(Y = y) = \sum_x P(x, y)$. Hence, we have: $p_Y(1) = \frac{1}{4} + \frac{1}{8} + \frac{1}{12} + \frac{1}{16} = \frac{25}{48}, p_Y(2) = \frac{1}{8} + \frac{1}{12} + \frac{1}{16} = \frac{13}{48}, p_Y(3) = \frac{1}{12} + \frac{1}{16} = \frac{7}{48}, p_Y(4) = \frac{1}{16}$

c) The conditional probability mass function of x given y can be calculated as: $p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$.

With $Y = 2$, $p_Y(y) = \frac{13}{48}$ (according to part b.). We have the conditional PMF of X given $Y = 2$: $p_{X|Y}(1|2) = \frac{P(1,2)}{p_Y(y)} = \frac{0}{13/48} = 0$, $p_{X|Y}(2|2) = \frac{P(2,2)}{p_Y(y)} = \frac{1/8}{13/48} = \frac{6}{13}$, $p_{X|Y}(3|2) = \frac{P(3,2)}{p_Y(y)} = \frac{1/12}{13/48} = \frac{4}{13}$ and $p_{X|Y}(4|2) = \frac{P(4,2)}{p_Y(y)} = \frac{1/16}{13/48} = \frac{3}{13}$

d) The conditional mean of X given $Y = 2$ or 3 is calculated as:

$$\begin{aligned}
\mathbb{E}[X|Y = 2 \text{ or } 3] &= \sum_{Y=\{2,3\}} x \cdot p_{X|Y}(x|y)\\
&= 1(p_{X|Y}(1|2) + p_{X|Y}(1|3)) + 2(p_{X|Y}(2|2) + p_{X|Y}(2|3))\\
&+ 3(p_{X|Y}(3|2) + p_{X|Y}(3|3)) + 4(p_{X|Y}(4|2) + p_{X|Y}(4|3))\\
&= 1 \cdot 0 + 2 \cdot \frac{6}{13} + 3 \cdot (\frac{4}{13} + \frac{4}{7}) + 4 \cdot (\frac{3}{13} + \frac{3}{7})\\
&= \frac{564}{91}
\end{aligned}$$

The conditional variance of X given $Y = 2, 3$ is calculated as:

$$\begin{aligned}
V[X|Y = 2 \text{ or } 3] &= \sum_{Y=\{2,3\}} (x - \mathbb{E}[X|Y])^2 \cdot p_{X|Y}(x|y)\\
&= (1 - \frac{564}{91})^2 \cdot 0 + (2 - \frac{564}{91})^2 \cdot \frac{6}{13} + (3 - \frac{564}{91})^2 \cdot (\frac{4}{13} + \frac{4}{7}) + (4 - \frac{564}{91})^2 \cdot (\frac{3}{13} + \frac{3}{7})\\
&= \frac{264}{13}
\end{aligned}$$

e) Let Z be the cost of each book where Z is an independent random variable). We know that $\mathbb{E}(Z) = 3$. $\mathbb{Y}$ is the expected number of books she will buy for each visit.

First, we need to calculate $\mathbb{E}(Y)$:

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y\in\Omega_Y} y \cdot p_Y(y)\\
&= 1(p_Y(1) + 2(p_Y(2)) + 3(p_Y(3)) + 4(p_Y(4))\\
&= \frac{25}{48} + 2\frac{13}{48} + 3\frac{7}{48} + 4\frac{1}{16}\\
&= \frac{7}{4}
\end{aligned}$$

The total expected expenditure of Mary's visit to the bookstore can be calculated as:

$$\text{Expected Expenditure} = \mathbb{E}(Z) * \mathbb{E}(Y) = 3 \cdot \frac{7}{4} = \frac{21}{4}$$

Problem 4. (15 points) Let Y_0 and Y_1 be two continuous random variables and $Z \sim \text{Bernoulli}(\alpha)$. Let X be a random variable defined as $X = ZY_1 + (1 - Z)Y_0$. Assuming that the probability density functions of Y_0 and Y_1 exist, show that the density of X is a mixture of the densities of Y_1 and Y_0 with α and $1 - \alpha$ as the mixing proportions, respectively.

Answer. The formula for the Bernoulli distribution is:

$$p(\omega) = \begin{cases} \alpha, & \text{if } \omega = S \\ 1 - \alpha, & \text{if } \omega = F \end{cases}$$

We know that the pdfs of Y_0 and Y_1 exists, or $f_{Y_0}(y_0)$ and $f_{Y_1}(y_1)$ exists. Hence,

$$\begin{aligned}
F_X(x) &= P((1 - Z)Y_0 + ZY_1 \leq x) = \int_{-\infty}^{x} f_X(x)dx = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x - ZY_1} f_{Y_0Y_1}((1 - z)y_0, zy_1)dy_0 \right) dy_1 \\
f_X(x) &= \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} \left(\frac{d}{dx} \int_{-\infty}^{x - ZY_1} f_{Y_0Y_1}((1 - z)y_0, zy_1)dy_0 \right) dy_1 \\
f_X(x) &= \int_{-\infty}^{\infty} f_{Y_0Y_1}(x - zy_1, zy_1)dy_1 = \int_{-\infty}^{\infty} f_{Y_0}(x - zy_1) f_{Y_1}(zy_1)dy_1 \\
&= f_{Y_0}(x - zy_1) \int_{-\infty}^{\infty} f_{Y_1}(zy_1)dy_1 - \int_{-\infty}^{\infty} f_{Y_1}(zy_1) \frac{d}{dy_1} f_{Y_0}(x - zy_1)dy_1 \\
f_X(x) &= (1 - z) f_{Y_0}(y_0) + z f_{Y_1}(y_1) = (1 - \alpha) f_{Y_0}(y_0) + \alpha f_{Y_1}(y_1)
\end{aligned}$$

Hence, the density of X is a mixture of density of Y_1 and Y_0 with α and $1 - \alpha$ as the mixing proportions.

Problem 5. (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 72 accidents over the next 8 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.

Answer.

a) We know that the formula for the maximum likelihood estimate of λ is: $\lambda_{ML} = \arg\max\{p(D|\lambda)\}$

   The formula for Poisson mass distribution is: $p(x|\lambda) = \frac{\lambda^x e^\lambda}{x!}$

Hence, we have:

$$\begin{aligned} p(D|\lambda) &= p(\{x_i\}_{i=1}^n|\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \prod_{i=1}^n \frac{\lambda^{x_i}e^-\lambda}{x_i!} \\ &= \frac{\lambda^{\sum x_i}e^{-n\lambda}}{\prod_{i=1}^n x_i!} \end{aligned}$$

Take the log-likelihood of the equation above:

$$ll(D,\lambda) = \ln\lambda\sum_{i=1}^n x_i - n\lambda - \sum_{i=1}^n (x_i)!$$

We take the derivative of the log-likelihood and set it to 0 for optimization:

$$\begin{aligned} \frac{\partial ll(D,\lambda)}{\partial\lambda} &= \frac{1}{\lambda}\sum_{i=1}^n x_i - n \\ &= 0 \end{aligned}$$

Hence, $\lambda_{ML} = \frac{\sum x_i}{n} = \frac{72}{8} = 9$.

b) We know that the formula for the maximum a posteriori estimate of λ is: $\lambda_{MAP} = \arg\max\{p(D|\lambda)(\lambda)\}$
The likelihood is: $p(D|\lambda) = \frac{\lambda^{\sum x_i}e^{-n\lambda}}{\prod_{i=1}^n x_i!}$
The prior density is: $p(\lambda) = \theta e^{-\theta\lambda}$
Log-likelihood:

$$\ln(\lambda|D) \propto \ln(D|\lambda) + \ln(\lambda) = \ln(\lambda)\sum_{i=1}^n x_i - n\lambda - \sum_{i=1}^n (x_i!) + \ln(\theta) - \lambda\theta$$

Optimization:

$$\lambda_{MAP} = \frac{\sum_{i=1}^n x_i}{\theta + n} = \frac{72}{8+0.5} = 8.47$$

c) From part b, we already know that:

$$p(D|\lambda) = \frac{\lambda^{\sum x_i}e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

and

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

Since we want to find the value of $\mathbb{E}[\Lambda|\mathcal{D}] = \int_0^\infty \lambda p(\lambda|\mathcal{D})d\lambda$, we need to find $p(\lambda|\mathcal{D})$.

$$\begin{aligned} p(\lambda|\mathcal{D}) &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{\int_0^\infty p(\mathcal{D}|\lambda)p(\lambda)d\lambda} \end{aligned}$$

We have:

$$\begin{aligned} p(\mathcal{D}) &= \int_0^\infty p(\mathcal{D}|\lambda)p(\lambda)d\lambda \\ &= \int_0^\infty \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \theta e^{-\theta\lambda} d\lambda \\ &= \int_0^\infty \frac{\theta\lambda^{\sum x_i} e^{-(n+\theta)\lambda}}{\prod_{i=1}^n x_i!} d\lambda \\ &= \frac{\theta\Gamma(\sum x_i + 1)}{\prod_{i=1}^n x_i!(n+\theta)^{\sum x_i + 1}} \end{aligned}$$

and subsequently:

$$\begin{aligned} p(\lambda|\mathcal{D}) &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})} \\ &= \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \cdot \theta e^{-\theta\lambda} \cdot \frac{\prod_{i=1}^n x_i!(n+\theta)^{\sum x_i + 1}}{\theta\Gamma(\sum x_i + 1)} \\ &= \frac{\lambda^{\sum x_i} e^{-(n+\theta)\lambda}(n+\theta)^{\sum x_i + 1}}{\Gamma(\sum x_i + 1)} \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}[\Lambda|\mathcal{D}] &= \int_0^\infty \lambda p(\lambda|\mathcal{D})d\lambda \\ &= \frac{\sum x_i + 1}{n+\theta} \\ &= \frac{73}{8.5} = 8.59 \end{aligned}$$

Problem 6. (15 points) Let X_1, X_2, …, X_n be i.i.d. Gaussian random variables, each having an unknown mean θ and known variance σ_0^2. If θ is itself selected from a normal population having a known mean μ and a known variance σ^2, determine

a) (5 points) the maximum a posteriori estimate of θ

b) (10 points) the Bayes estimate of θ.

Hint: look into conjugate priors for the Gaussian distribution. Chapter 2 of Bishop's textbook will be useful as well as resources on the internet.

Answer.

a) We know that the formula for the maximum a posteriori estimate of λ is: $\lambda_{MAP} = \arg\max\{p(D|\theta)p(\theta)\}$.

The likelihood is: $p(\mathcal{D}|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(\sum X_i - \theta)^2}$

The conjugate prior of the mean is: $p(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\theta-\mu}{\sigma})^2}$

Log-likelihood:

$$\ln(\theta|D) \propto \ln(D|\theta) + \ln(\theta) = \frac{1}{2}\ln(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(\sum X_i - \theta)^2 - \ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}(\frac{\theta-\mu}{\sigma})^2$$

Optimization:

$$-\frac{1}{\sigma_0^2}\theta - \frac{1}{\sigma_0^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\theta - \frac{1}{\sigma^2}\mu = 0$$

$$\theta_{MAP} = \frac{\frac{1}{\sigma_0^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\mu}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}$$

b) We have:

$$\begin{aligned} p(\theta) &= \int_0^\infty p(\mathcal{D}|\theta)p(\theta)d\theta \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(\sum X_i - \theta)^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\theta-\mu}{\sigma})^2} d\theta \\ &= \int_0^\infty \frac{e^{-\frac{1}{2\sigma_0^2}(\sum X_i - \theta)^2 - \frac{1}{2}(\frac{\theta-\mu}{\sigma})^2}}{2\pi\sigma\sigma_0} d\theta \end{aligned}$$

Problem 7. (40 points) Expectation-maximization (EM) algorithm. Let X be a random variable distributed according to $p_X(x)$ and Y be a random variable distributed according to $p_Y(y)$. Let $\mathcal{D}_X = \{x_i\}_{i=1}^m$ be an i.i.d sample from $p_X(x)$ and $\mathcal{D}_Y = \{y_i\}_{i=1}^n$ be an i.i.d. sample from $p_Y(y)$. Finally, let $p_X(x)$ and $p_Y(y)$ be defined as follows

$$p_X(x) = \alpha\mathcal{N}(\mu_1, \sigma_1^2) + (1-\alpha)\mathcal{N}(\mu_2, \sigma_2^2)$$

and

$$p_Y(y) = \beta\mathcal{N}(\mu_1, \sigma_1^2) + (1-\beta)\mathcal{N}(\mu_2, \sigma_2^2),$$

where $\mathcal{N}(\mu, \sigma^2)$ is a univariate Gaussian distribution with mean μ and variance σ^2, $\alpha \in (0,1)$, $\beta \in (0,1)$, $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, $\sigma_1 \in \mathbb{R}^+$ and $\sigma_2 \in \mathbb{R}^+$ are unknown parameters.

a) (5 points) Derive update rules of the EM algorithm for estimating α, μ_1, μ_2, σ_1, and σ_2 based only on data set $\mathcal{D}_X$.

b) (15 points) Derive update rules of an EM algorithm for estimating α, β, μ_1, μ_2, σ_1, and σ_2 from data sets $\mathcal{D}_X$ and $\mathcal{D}_Y$.

c) (20 points) Implement both learning algorithms from above and evaluate them on simulated data when $m, n = 100$ and $m, n = 1000$. However, in each case repeat the experiment $B = 1000$ times to estimate the expectation and variance of all parameters. To do so, repeatedly draw samples $\mathcal{D}_X$ and $\mathcal{D}_Y$ and then estimate the parameters based on these samples. Finally, average those B estimates and calculate their mean and variance. Document all experiments and discuss your findings.

To implement the EM algorithm you must make some practical decisions on how to stop the estimation process. You may decide to impose the maximum number of steps, stop the algorithm when the updated parameters stabilize, stop the algorithm when the log-likelihood stabilizes, or some combination thereof. In either case, experiment first with the stoppage criterion and once it is fixed carry out the experiment.

Answer.

a) We know that $p_X(x)$ is a mixture of 2 Guassian distribution with the mixing weights of α and $1-\alpha$.

Let z be a random variable with the property of 1-of-2 presentation in which only one of z_k ($k \in \{1,2\}$) has the value of 1 and the other equal 0. The marginal distribution of z is: $p(z_1 = 1) = \alpha$ and $p(z_2 = 1) = 1 - \alpha$.

Let $x_i \in \mathcal{D}_X$. The conditional probability of z given x (at step t) is:

$$\gamma(z_{i1}) = \frac{\alpha^{t-1} \cdot \mathcal{N}(x_i | \mu_1^{t-1}, \sigma_1^{t-1})}{\alpha^{t-1} \cdot \mathcal{N}(x_i | \mu_1^{t-1}, \sigma^{t-1}) + (1 - \alpha^{t-1}) \cdot \mathcal{N}(x_i | \mu_2^{t-1}, \sigma_2^{t-1})}$$

$$\gamma(z_{i2}) = \frac{(1 - \alpha^{t-1}) \cdot \mathcal{N}(x_i | \mu_2^{t-1}, \sigma_2^{t-1})}{\alpha^{t-1} \cdot \mathcal{N}(x_i | \mu_1^{t-1}, \sigma_1^{t-1}) + (1 - \alpha^{t-1}) \cdot \mathcal{N}(x_i | \mu_2^{t-1}, \sigma_2^{t-1})}$$

Define:

$$N_1 = \sum_{i=1}^{m} \gamma(z_{i1})$$

$$N_2 = \sum_{i=1}^{m} \gamma(z_{i2})$$

Update rules for parameters α, μ_1, μ_2, σ_1, and σ_2 at step t:

$$\mu_1^t = \frac{1}{N_1} \sum_{i=1}^{m} \gamma(z_{i1}) x_i$$

$$\mu_2^t = \frac{1}{N_2} \sum_{i=1}^{m} \gamma(z_{i2}) x_i$$

$$\sigma_1^t = \frac{1}{N_1} \sum_{i=1}^{m} \gamma(z_{i1})(x_i - \mu_1^t)(x_i - \mu_1^t)^T$$

$$\sigma_2^t = \frac{1}{N_2} \sum_{i=1}^{m} \gamma(z_{i2})(x_i - \mu_2^t)(x_i - \mu_2^t)^T$$

$$\alpha^t = \frac{N_1}{m}$$

b) For each point $x_i \in \mathcal{D}_X$, the update rule for EM algorithm works similar to part a. For each point $y_i \in \mathcal{D}_Y$, the conditional probability of z' given y_i (at step t) is

$$\gamma(z'_{i1}) = \frac{\beta^{t-1} \cdot \mathcal{N}(y_i | \mu_1^{t-1}, \sigma_1^{t-1})}{\beta^{t-1} \cdot \mathcal{N}(y_i | \mu_1^{t-1}, \sigma^{t-1}) + (1 - \beta^{t-1}) \cdot \mathcal{N}(y_i | \mu_2^{t-1}, \sigma_2^{t-1})}$$

$$\gamma(z'_{i2}) = \frac{(1 - \beta^{t-1}) \cdot \mathcal{N}(y_i | \mu_2^{t-1}, \sigma_2^{t-1})}{\beta^{t-1} \cdot \mathcal{N}(y_i | \mu_1^{t-1}, \sigma_1^{t-1}) + (1 - \beta^{t-1}) \cdot \mathcal{N}(y_i | \mu_2^{t-1}, \sigma_2^{t-1})}$$

Define:

$$N_3 = \sum_{i=1}^{n} \gamma(z' i1)$$

$$N_4 = \sum_{i=1}^{n} \gamma(z' i2)$$

Variables	True value	Expectation	Variance
α	0.1	0.29	0.05
μ_1	5.0	-0.52	1.41
μ_2	-1.0	-0.005	7.41
σ_1	1.0	2.94	3.65
σ_2	1.3	2.2	3.83

Table 1: Algorithm 1, $m, n = 100$

Variables	True value	Expectation	Variance
α	0.1	0.28	0.05
μ_1	5.0	-0.048	3.68
μ_2	-1.0	-0.52	3.91
σ_1	1.0	3.049	4.75
σ_2	1.3	2.3	5.11

Table 2: Algorithm 1, $m, n = 1000$

Update rules for parameters β, μ_1, μ_2, σ_1, and σ_2 at step t:

$$\mu_1^t = \frac{1}{N_3}\sum_{i=1}^{n}\gamma(z'_{i1})y_i$$

$$\mu_2^t = \frac{1}{N_4}\sum_{i=1}^{n}\gamma(z'_{i2})y_i$$

$$\sigma_1^t = \frac{1}{N_3}\sum_{i=1}^{n}\gamma(z'_{i1})(y_i - \mu_1^t)(y_i - \mu_1^t)^T$$

$$\sigma_2^t = \frac{1}{N_4}\sum_{i=1}^{n}\gamma(z'_{i2})(y_i - \mu_2^t)(y_i - \mu_2^t)^T$$

$$\beta^t = \frac{N_3}{n}$$

c) The tables below show the results of running the four experiment.

The data is sampled from 2 univariate Guassian distribution, the first one has $\mu_1 = 5.0$, $\sigma_1 = 1.0$ and the second one has $\mu_2 = -1.0$, $\sigma_2 = 1.3$. I set $\alpha = 0.1$ and $\beta = 0.8$.

The first and second experiments are for the algorithm in part (a), with $m, n = 100$ and $m, n = 1000$ respectively.

Problem 8. (30 points) Properties of high-dimensional spaces.

a) (10 points) Show that in a high-dimensional space, most of the volume of a hypercube is concentrated in corners, which themselves become very long "spikes." Hint: compute the ratio of the volume of a hypersphere with radius r to the volume of a hypercube with side $2r$ around it and also the ratio of the distance from the center of the hypercube to one of the corners divided by the distance to one of the sides.

b) (10 points) Show that for points uniformly distributed inside a sphere in d dimensions, where d is large, almost all of the points are concentrated in a thin shell close to the surface. Hint: compute the

Variables	True value	Expectation	Variance
α	0.1	0.33	0.08
β	0.8	0.48	0.1
μ_1	5.0	0.46	4.97
μ_2	-1.0	1.83	9.26
σ_1	1.0	2.54	2.83
σ_2	1.3	1.45	1.20

Table 3: Algorithm 2, $m, n = 100$

Variables	True value	Expectation	Variance
α	0.1	0.25	0.007
β	0.8	0.49	0.13
μ_1	5.0	0.09	4.67
μ_2	-1.0	-1.2	10.09
σ_1	1.0	2.69	2.72
σ_2	1.3	1.85	3.67

Table 4: Algorithm 2, $m, n = 1000$

fraction of the volume of the sphere which lies at the distance between $r - \epsilon$ and r from the center, where $0 < \epsilon < r$. Evaluate this fraction for $\epsilon = 0.01r$ and also for $\epsilon = 0.5r$ for $d \in \{1, 2, 3, 10, 100\}$.

c) (10 points) Evaluate computationally what you derived. First, generate n d-dimensional data points uniformly at random within a hypercube with side $2r$. Then, compute the fraction of points f that are within the hypersphere of radius r inscribed in the hypercube. Do this for d ranging from 1 to 100. Generate the plot of f as a function of d (make sure axes are labeled). Pick n to be at least 100, but a larger number is desirable, depending on your computational resources and also mark the values computed from the formula derived in part (a). If you pick a relatively small n, and have computational resources, visualize the uncertainty of the estimated fraction over multiple trials for the same d. Evaluate also what happens if you vary r for a fixed d? Describe what you see in all plots and give your reasoning as to why it happened.

Answer.

a) Consider a hypercube with dimension n with side $2r$ and an n-dimensional sphere inscribed within it.

The volume of the hypercube is:

$$V(C) = (2r)^n$$

The volume of the sphere is:

$$V(S) = r^n \cdot \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)}$$

The ratio of the two volume is:

$$\frac{V(S)}{V(C)} = \frac{r^n \cdot \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}}{(2r)^n}$$

We want to see how this ratio change as we increase the number of dimensions:

$$\begin{aligned}\lim_{n\to\infty} \frac{r^n \cdot \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}}{(2r)^n} &= \lim_{n\to\infty} \frac{\cdot \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}}{2^n} \\ &= \lim_{n\to\infty} (\frac{\sqrt{\pi}}{2})^n \frac{1}{\Gamma(\frac{n}{2}+1)} \\ &= 0\end{aligned}$$

Since the ratio goes to 0 as the dimensions increase, we see that most of the volume of the hypercube is not in the center, but in the corner of the n-cube.

The ratio of the distant from the center of the hypercube to one of the corners over the distance to one of the side is:

$$\begin{aligned}\frac{\text{center to one to the corner}}{\text{center to one of the side}} &= \frac{\sqrt{r^2 \cdot n}}{r} \\ &= \sqrt{n} \\ \lim_{n\to\infty} \sqrt{n} = +\infty\end{aligned}$$

Hence, we see that as n increases, the corners will become very long spikes.

b) The Volume of the d-dimension sphere with radius r is given by the formula:

$$V_d(r) = r^n \cdot \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})}$$

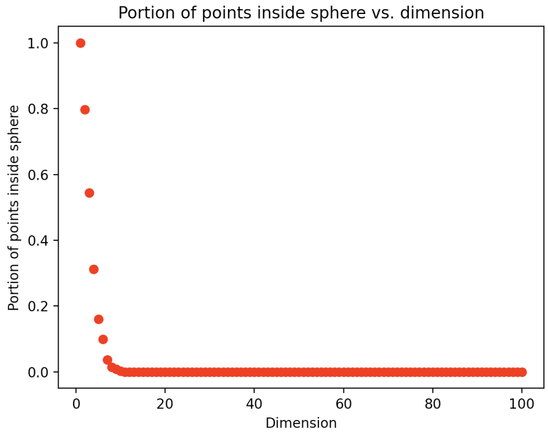
Let T be the volume of the portion that lies between radius r and $r-\epsilon$ from the center of the sphere:

$$\begin{aligned}V(T) &= \frac{V_d(r) - V_d(r-\epsilon)}{V_d(r)} \\ &= \frac{r^n \cdot \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})} - (r-\epsilon)^n \cdot \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})}}{r^n \cdot \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})}} \\ &= \frac{r^n - (r-\epsilon)^n}{r^n}\end{aligned}$$

As n increases, this fraction goes to 1. Hence, most of the points in the sphere are concentrated near the surface.

ϵ	d	$V(T)$
$0.01r$	1	0.01
$0.01r$	2	0.0199
$0.01r$	3	0.029701
$0.01r$	10	0.09561
$0.01r$	100	0.6339
$0.001r$	1	0.001
$0.001r$	2	0.00199
$0.001r$	3	0.002997
$0.001r$	10	0.0099
$0.001r$	100	0.0952

Figure 1: Portion of points inside the sphere as d increases for $r = 1$



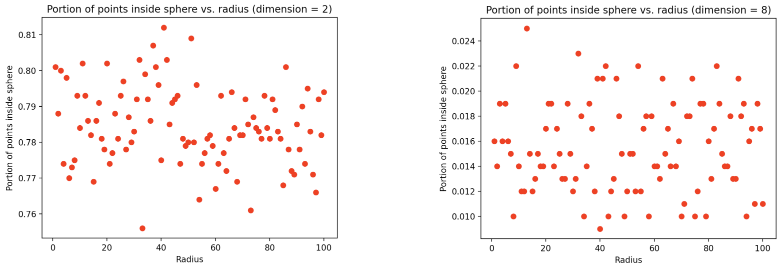


Figure 2: Effects of varying r for $d = 2$ and $d = 8$

c) Figure 1 shows the plots for the simulation of the formula in part (a). In the experiment, I have set the number of points to be 1000, the dimension d ranging from 1 to 100, and the radius $r = 1$. As we can see from the graph, as $d \to \infty$, the portion of points inside the sphere goes to 0.

Figure 2 shows the effects of varying r on our function. As we can see for the two scenario of $d = 2$ and $d = 8$, r ranging from 1 to 100, even though the portion of points does not remain constant since we are doing random simulation, all the points fall closely to the value computed by the close-formed formula in part (a).