Junda An

Homework 1

CSCI 420
February 26, 2019

AML EX:1.[1-6, 10]

Answer to the problem goes here.

1. EX 1.1 answer here.

   (a) Input space: set of all possible patients' symptoms.

   Output space: set of all possible patients' illness.

   Target function: ideal formula for patients' illness based on patients' symptoms.

   Specifics of the data set: medical histories of patients.

   (b) Input space: set of all possible handwritten postal zip codes.

   Output space: set of all possible actual postal zip codes.

   Target function: ideal formula for figuring out actual postal zip codes based on handwritten postal zip codes.

   Specifics of the data set: historical documentations of handwritten postal zip codes and their corresponding actual postal zip codes.

   (c) Input space: a set of all possible features of the email, like its subject and contents.

   Output space: a set of all possible output, here yes or no.

   Target function: ideal formula for determining whether an email is spam based on features of the email.

   Specifics of the data set: historical records of all emails.

   (d) Input space: a set of all possible prices, temperatures, and days of the week.

   Output space: electric load

   Target function: ideal formula for determining electric load based on price, temperature and day of the week.

   Specifics of the data set: historical records of price, temperature, and day of the week and the corresponding electric load.

   (e) Input space: a set of all possible data

   Output space: a set of all possible empirical solutions

   Target function: ideal formula for determining the empirical solution based on data

   Specifics of the data set: data

Homework 1

2. Ex 1.2 answer here.

   (a) "For free", "pure profits", "apply online", and "the best rates" are some keywords that will end up with a large positive weight in the perceptron.

   (b) "Dear Professor", "Sincerely", college names like "College of William and Mary" are some keywords that will get a negative weight.

   (c) The bias value directly affects how many borderline messages end up being classified as spam.

3. Ex 1.3 answer here.

   (a) Suppose $y(t) < 0$. Since $x(t)$ is misclassified by $w(t)$ and $y(t) < 0$, $w^T(t)x(t) > 0$. Therefore, $y(t)w^T(t)x(t) < 0$ Suppose $y(t) > 0$. Since $x(t)$ is misclassified by $w(t)$ and $y(t) > 0$, $w^T(t)x(t) < 0$. Therefore, $y(t)w^T(t)x(t) < 0$.

   (b) $y(t)w^T(t+1)x(t) = y(t)(w(t) + y(t)x(t))^T x(t)$(since (1.3)).

   $= y(t)(w^T(t) + x^T(t)y(t))x(t)$ (since $y(t)$ is a scalar)

   $= y(t)w^T(t)x(t) + y(t)x^T(t)y(t)x(t)$

   $= y(t)w^T(t)x(t) + y^2(t)(x^T(t)x(t))$

   $= y(t)w^T(t)x(t) + y^2(t)||x(t)||^2 > y(t)w^T(t)x(t)$ (Since $y^2(t)||x(t)||^2 > 0$ where $y(t)$ is nonzero and $x(t)$ is a nonzero vector)

   (c) Let $(x(t), y(t))$ be a point that we concern. If it is classified correctly, we do not update weights. If it is incorrectly classified, according to the updating rule, $w(t+1) = w(t) + y(t)x(t)$, where $w(t+1)$ is the updated weights and $w(t)$ is the original weights.

   Then, $w^T(t+1)x(t) = (w(t)+y(t)x(t))^T x(t) = w^T(t)x(t)+y(t)x^T(t)x(t)$ (take transpose and distribute x(t)) $= w^T(t)x(t) + y(t)||x(t)||^2$.

   If $y(t) = 1$, since $(x(t), y(t))$ is incorrectly classified, $w^T(t)x(t) < 0$. From the equation above, $w^T(t+1)x(t) = w^T(t)x(t) + ||x(t)||^2$. Since $||x(t)||^2 > 0$, $w^T(t+1)x(t) > w^T(t)x(t)$. Therefore, the move from $W(t+1)$ to $w(t)$ is moving in a right direction. (The direction is towards positive).

   If $y(t) = -1$, since $(x(t), y(t))$ is incorrectly classified, $w^T(t)x(t) > 0$. From the equation above, $w^T(t+1)x(t) = w^T(t)x(t) - ||x(t)||^2$. Since $||x(t)||^2 > 0$, $w^T(t+1)x(t) < w^T(t)x(t)$. Therefore, the move from $W(t+1)$ to $w(t)$ is moving in a right direction.(The direction is towards negative)

4. Ex 1.4 answer here.

   Please see the jupyter notebook. It usually takes less than 15 iterations to run. The final hypothesis decision boundary is pretty closed to the target decision boundary.

5. Ex 1.5 answer here

(a), (c), (e) are more suitable for learning approach, because these problems are less specified and one needs data to pin down the target function.

(b), (d) are more suitable for design approach, because people can analytically derive the target function to classify numbers into primes and non-primes, and they can also derive the target function to determine the time it would take a falling object to hit the ground.

6. Ex 1.6 answer here

   (a) can fit both supervised learning and unsupervised. For supervised learning, the system uses data of ratings of books from current users and recommends the book with highest rating to the user. For unsupervised learning, the system uses data on different features of books that the user like the most and recommends a book that has these features.

   (b) fits reinforcement learning. The input space should be a particular tic tac toe situation. The output space should be the next move chosen by the player. The system can report how well the next move is and use these data as training data. It then can find the best strategy to play the game based on these training data.

   (c) can fit both supervised learning and unsupervised learning. For supervised learning, the system uses data on different features of movies and the movies' corresponding categories. It can categorize movies into different types based on the final hypothesis function on determining types of movies. So, the output space should be different types of movies. For unsupervised learning, the system uses data uses data on different features of movies and categories movies into different types based on these features without knowing the specific types of these movies.

   (d) fits reinforcement learning. The input space can be ryhthms, melodies or other elements that describe the style of music. The output should be music produced. The system reinforces on how listeners think of the music, like good or bad.

   (e) fits supervised learning. The system uses data of information of current customers, like their income, ages, debts, and so on, to get a final hypothesis function. Then, for each bank customer, the system take his information and determine the credit limit, based on the function.

7. Ex 1.10 answer here

   (a) Since coins are fair coins, $\mu = 0.5$

   (b) (c) see jupyter notebook named Ex1.10

   (d) c_1 and c_{rand} obey Hoeffding bound, because they only consider one bin at a time. However, c_{min} does not obey Hoeffding bound, since it considers multiple bins and choose the best one (the one with minimum heads).

PRML 1.[1,5,6,9,11,12]

1. EX 1.1 answer here.

   To find coefficients $\{w_i\}$ that minimizes the error function, we need to take derivative of the error function with respect to each w_i and find w_i that makes the derivative zero.

$$\frac{\partial \frac{\sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)^2}{2}}{\partial w_i}$$

$$\Rightarrow \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)x_n^i = 0$$

$$\Rightarrow \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j)x_n^i = \sum_{n=1}^{N} t_n x_n^i$$

$$\Rightarrow \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^{i+j}) = \sum_{n=1}^{N} t_n x_n^i$$

   (put x_n^i into the inner summation, since it is independent of j)

$$\Rightarrow \sum_{j=0}^{M}(\sum_{n=1}^{N} w_j x_n^{i+j}) = \sum_{n=1}^{N} t_n x_n^i$$

   (exchange the summation symbols)

$$\Rightarrow \sum_{j=0}^{M}(\sum_{n=1}^{N} x_n^{i+j}) w_j = \sum_{n=1}^{N} x_n^i t_n$$

   (get w_j out of inner summation, since it is independent of n)

   Let $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$, $T_i = \sum_{n=1}^{N} x_n^i t_n$. Therefore, $\sum_{j=0}^{M} A_{ij} w_j = T_i$

2. Ex 1.5 answer here

$$\begin{aligned}
var(f) &= E[(f(x) - E[f(x)])^2] \text{ (1.38)} \\
&= E[f(x)^2 - 2f(x)E[f(x)] + E[f(x)]^2] \text{ (\textit{expanding the square})} \\
&= E[f(x)^2] - E[2f(x)E[f(x)]] + E[E[f(x)]^2] \text{ (\textit{distributing E})} \\
&= E[f(x)^2] - 2E[f(x)]E[f(x)] + E[f(x)]^2 \text{ (\textit{since } } E[f(x)]^2 \text{ \textit{is a constant})} \\
&= E[f(x)^2] - E[f(x)]^2
\end{aligned}$$

3. Ex 1.6 answer here

   For the discrete case, let $p(x), p(y)$ be probability mass function for x and y. Since x, y are independent, $p(x, y) = p(x)p(y)$. Then,

$$\begin{aligned} E_{x,y}[x, y] &= \sum_x \sum_y (p(x, y)xy) \ (1.33) \\ &= \sum_x \sum_y (p(x)p(y)(x)(y)) \ (\textit{since } x \textit{ and } y \textit{ are independent}) \\ &= (\sum_x p(x)x)(\sum_y p(y)y) \\ &(\textit{since } p(x)x \textit{ is independent of } y, \textit{ and } p(y)y \textit{ is independent of } x) \\ &= E[x]E[y] \ (1.33) \end{aligned}$$

   For the continuous case, let $p(x), p(y)$ be probability density function for x and y. Since x, y are independent, $p(x, y) = p(x)p(y)$. Then,

$$\begin{aligned} E_{x,y}[x, y] &= \int \int p(x, y)xydxdy \ (1.34) \\ &= \int \int p(x)p(y)(x)(y)dxdy \ (\textit{since } x \textit{ and } y \textit{ are independent}) \\ &= (\int p(x)xdx) \int p(y)ydy \\ &(\textit{since } p(x)x \textit{ is independent of } y, \textit{ and } p(y)y \textit{ is independent of } x) \\ &= E[x]E[y] \ (1.34) \end{aligned}$$

   Therefore,

$$\begin{aligned} cov[x, y] &= E_{x,y}[x, y] - E[x]E[y] \ (\textit{since } 1.42) \\ &= E[x]E[y] - E[x]E[y] \\ &= 0 \end{aligned}$$

   Hence, if two variables x and y are independent, then their covariance is zero.

4. Ex 1.9 answer here

   (1.46): To find the maximum, we need to differentiate (1.46) with respect to x.

$$\frac{\partial \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\partial x} = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} (-\frac{1}{2\sigma^2}) 2(x - \mu) = 0$$

   Therefore, $x = \mu$ is the only root of the partial differential equation. Hence, the mode of the Gaussian distribution (1.46) is given by .

(1.52): To find the maximum, we need to differentiate (1.52) with respect to x.
$\frac{\partial \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}{\partial x} = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} (-\frac{1}{2})(\Sigma^{-1} + (\Sigma^{-1})^T)(x-\mu)$
(since $\frac{\partial x^T A x}{\partial x} = (A + A^T)x$)
Therefore, $x = \mu$ is the only root of the partial differential equation.

5. Ex 1.11 answer here

$$\frac{\partial ln p(x|\mu, \sigma^2)}{\partial \mu} = 0$$

$$\frac{\partial -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} ln\sigma^2 - \frac{N}{2} ln(2\pi)}{\partial \mu} = 0 \ (1.54)$$

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (-2)(x_n - \mu) = 0$$

$$\Rightarrow \frac{1}{2} (\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} \mu) = 0$$

$$\Rightarrow \frac{1}{2} (\sum_{n=1}^{N} x_n - N\mu) = 0$$

Therefore, $\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$ is the only root of the partial differential equation. Therefore, $\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$ is the maximum likelihood solution for the mean.

$$\frac{\partial ln p(x|\mu, \sigma^2)}{\partial \sigma^2} = 0$$

$$\frac{\partial -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} ln\sigma^2 - \frac{N}{2} ln(2\pi)}{\partial \sigma^2} = 0 \ (1.54)$$

$$\Rightarrow \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2\sigma^2} = 0$$

$$\Rightarrow \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 = \frac{N}{2\sigma^2}$$

$$\Rightarrow \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 = \sigma^2$$

Therefore, $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$ is the only root of the partial differential equation. Hence, $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$ is the maximum likelihood solution for the variance.

6. Ex 1.12 answer here

   If $n = m$, $E[x_n x_m] = E[x_n^2] = \mu^2 + \sigma^2 = \mu^2 + I_{nm}\sigma^2$, where $Inm = 1$ (according to 1.50)

   If $n \neq m$, $E[x_n x_m] = E[x_n]E[x_m] = \mu\mu = \mu^2 = \mu^2 + I_{nm}\sigma^2$, where $I_{nm} = 0$ (since x_m and x_n are two independent data points sampled from Gaussian distribution with mean μ.)

$$\begin{aligned}
E[\mu_{ML}] &= E[\frac{1}{N}\sum_{n=1}^{N} x_n] \ (plug\ in\ (1.55)) \\
&= \frac{1}{N}\sum_{n=1}^{N} E[x_n] \ (put\ E\ into\ the\ summation) \\
&= \frac{1}{N}\sum_{n=1}^{N} \mu \ (E[x_n] = \mu) \\
&= \frac{1}{N} N\mu \ (since\ \mu\ is\ a\ constant) \\
&= \mu
\end{aligned}$$

$$\begin{aligned}
E[\sigma^2_{ML}] &= E[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2] \ (plug\ in\ (1.56)) \\
&= \frac{1}{N}\sum_{n=1}^{N} E[(x_n - \mu_{ML})^2] \ (put\ E\ into\ the\ summation) \\
&= \frac{1}{N}\sum_{n=1}^{N} E[x_n^2 - 2x_n\mu_{ML} + \mu^2_{ML}] \ (expanding\ square) \\
&= \frac{1}{N}\sum_{n=1}^{N} (E[x_n^2] - 2E[x_n\mu_{ML}] + E[\mu^2_{ML}]) \ (A) \ (distributing\ E)
\end{aligned}$$

$E[x_n^2] = \mu^2 + \sigma^2$ (I) (according to (1.40)).

$$E[x_n \mu_{ML}] = E[x_n \frac{1}{N} \sum_{m=1}^{N} x_m] \ (substitute\ (1.55))$$
$$= \frac{1}{N} \sum_{m=1}^{N} E[x_n x_m] \ (put\ E\ into\ the\ summation)$$

Since $1 \leq n \leq N$ and m goes from 1 to N, $x_m = x_n$ exactly first time in the N iteration. Therefore, according to (1.130), $\frac{1}{N} \sum_{m=1}^{N} E[x_n x_m] = \frac{1}{N}((\mu^2 + \sigma^2) + (N-1)\mu^2) = \mu^2 + \frac{1}{N}\sigma^2$ (II).

Substitute (1.55) into $E[\mu_{ML}^2]$ and we get

$$E[\mu_{ML}^2] = E[(\frac{1}{N} \sum_{m=1}^{N} x_m)^2]$$
$$= \frac{1}{N^2} E[(\sum_{m=1}^{N} x_m)^2]$$
$$= \frac{1}{N^2} E[(x_1x_1 + x_1x_2 + ... + x_1x_N) + ... + (x_Nx_1 + x_Nx_2 + ... + x_Nx_N)]$$
$$(expanding\ the\ summation)$$
$$= \frac{1}{N^2} E[\sum_{m=1}^{N} \sum_{i=1}^{N} x_m x_i] \ (1 \leq m, i \leq N.\ m\ and\ i\ are\ indices\ of\ x)$$
$$= \frac{1}{N^2} \sum_{m=1}^{N} \sum_{i=1}^{N} E[x_m x_i] \ (put\ E\ into\ the\ summations)$$

For each iteration on m, there is N iterations on i, in which $x_m = x_i$ exactly once. Since there are N iterations on m, there are N^2 iterations and $x_m = x_i$ N times. Therefore, according to (1.130) $E[\mu_{ML}^2] = \frac{1}{N^2} \sum_{m=1}^{N} \sum_{i=1}^{N} E[x_m x_i] = \frac{1}{N^2}(N(\mu^2 + \sigma^2) + (N^2 - N)\mu^2) = \mu^2 + \frac{1}{N}\sigma^2$ (III).

Plug in (I), (II), (III) into (A), we can get

$$\begin{aligned}
E[\sigma^2_{ML}] &= \frac{1}{N}\sum_{n=1}^{N}(E[x_n^2] - 2E[x_n\mu_{ML}] + E[\mu^2_{ML}]) \ (A) \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mu^2 + \sigma^2 - 2(\mu^2 + \frac{1}{N}\sigma^2) + \mu^2 + \frac{1}{N}\sigma^2) \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{N-1}{N}\sigma^2 \\
&= \frac{1}{N}N\frac{N-1}{N}\sigma^2 \ (since \ \sigma^2 \ is \ a \ constant) \\
&= \frac{N-1}{N}\sigma^2
\end{aligned}$$