1: AML Pg. 4, Ex. 1.1

Answer:

a)
$\mathcal{X}$: Set of all possible permutations of medical histories and symptoms of patients
$\mathcal{Y}$: Set of all possible problems with patients
f: $\mathcal{X} \to \mathcal{Y}$: Ideal formula to predict problem based on medical history and symptoms.

b)
$\mathcal{X}$: Set of all possible permutations of pixel intensities in a picture of one or more handwritten digits.
$\mathcal{Y}$: The nine possible digits (In zip code example, take a picture of each digit individually – otherwise, $\mathcal{Y}$ is set of all permutations of 5 digits)
f: $\mathcal{X} \to \mathcal{Y}$: Ideal formula to predict digit(s) based on input picture of handwritten digit(s).

c)
$\mathcal{X}$: Set of all possible permutations of characters in an email and contents of files attached to emails.
$\mathcal{Y}$: Spam – true or false.
f: $\mathcal{X} \to \mathcal{Y}$: Ideal formula to predict if email is spam or not based on input email.

d)
$\mathcal{X}$: Set of all possible permutations of price, temperature, and day-of-the-week values. For most practical cases, price and temperature may be capped at specified upper- and lower-bounds (0 lower-bound for price).
$\mathcal{Y}$: The continuous variable electric load.
f: $\mathcal{X} \to \mathcal{Y}$: Ideal formula to predict electric load given input price, temperature, and day of the week.

e)
$\mathcal{X}$: Set of all possible permutations of pixel intensities in a satellite image of a Metropolitan Statistical Area (MSA).
$\mathcal{Y}$: Set of all possible (or economically reasonable – not negative or greater than 100) unemployment percentages for a given MSA.
f: $\mathcal{X} \to \mathcal{Y}$: Ideal formula to predict unemployment of MSA given a satellite image of it.

2: AML Pg. 6, Ex. 1.2

Answer:
a) winner, money, SSN, arrest, comply, hours, minutes, phone
b) thanks, best, regards, respectfully, hello, class, academics, confirmation
c) The bias directly affects how many borderline messages end up being classified as spam. If a borderline message is defined as one whose sum of weights times inputs is near 0, a highly positive bias will result in that message being classified as spam, and a highly negative bias will result in that message not being classified as spam.

3: AML Pg. 8, Ex. 1.3

Answer:

a)
$h(\mathbf{x(t)}) = \mathbf{w^T(t)x(t)}$
For misclassification, $y(t) = -h(\mathbf{x(t)})$.
Proof by cases:
Case 1:
If $y(t) = +1$, $\boldsymbol{w^T}(t)\boldsymbol{x}(t) = -1$, so $y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) = -1 < 0$
Case 2:
If $y(t) = -1$, $\boldsymbol{w^T}(t)\boldsymbol{x}(t) = +1$, so $y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) = -1 < 0$

b) By eq. 1.3:

$$\begin{aligned} y(t)\boldsymbol{w^T}(t+1)\boldsymbol{x}(t) &= y(t)[\boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)]^T\boldsymbol{x}(t) \\ &= y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) + y^2(t)\boldsymbol{x^T}(t)\boldsymbol{x}(t) \\ \therefore y(t)\boldsymbol{w^T}(t+1)\boldsymbol{x}(t) &> y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) \end{aligned}$$

holds if and only if

$$y^2(t)\boldsymbol{x^T}(t)\boldsymbol{x}(t) > 0$$

Both $y^2(t)$ and $\boldsymbol{x^T}(t)\boldsymbol{x}(t)$ are positive – the former because any scalar squared is positive, and the latter because $\boldsymbol{x^T}(t)\boldsymbol{x}(t)$ is a sum of the squared scalars within $\boldsymbol{x}(t)$. Both $y^2(t)\boldsymbol{x^T}(t)\boldsymbol{x}(t) > 0$ and subsequently $y(t)\boldsymbol{w^T}(t+1)\boldsymbol{x}(t) > y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t)$ are therefore true.

c) If $h(\boldsymbol{x}(t)) = -1$ and the point $\boldsymbol{x}(t)$ was misclassified, $y(t) = +1$. The update rule $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)$ will therefore make the new decision boundary more positive. That is, for any t, $\boldsymbol{w}^T(t+1)\boldsymbol{x}(t)$ is more likely to be positive than $\boldsymbol{w}^T(t)\boldsymbol{x}(t)$. A future point similar to $\boldsymbol{x}(t)$ is therefore less likely to be misclassified as negative. The same argument holds with $h(\boldsymbol{x}(t)) = +1$ and $y(t) = -1$, where in this case the decision boundary is made more negative. Each misclassified point causes an adjustment that makes nearby points less likely to be misclassified in the same way.

4: AML Pg. 8, Ex. 1.4

Answer: See code.

5: AML Pg. 11, Ex. 1.5

Answer:

a) Determining the age at which a particular medical test should be performed is a suitable application of the design approach if one can analytically derive the model relating a person's age to the probability of that person contracting a condition mandating the test. This model, however, seems beyond what biology can currently analytically derive. Learning would therefore be more suitable.

b) Constructing an algorithm to determine if a number is prime consists of a simple loop and divisions. This is therefore a problem more suited to design than learning.

c) No derivable model relates the probability of credit card fraud to the characteristics or activity of a person, especially given that potentially randomly varying perniciousness of fraudsters is also a factor. Learning is more suited to this problem.

d) Determining the time it would take a falling object to hit the ground can be done with a model designed by an Intro Physics student. The design approach is more suitable.

e) The optimal cycle for traffic lights in a busy intersection depends on traffic patterns, which can be highly random and fluctuate both daily, seasonally, and hourly. The extent of these fluctuations is not derivable from first principles. The learning approach is more suitable.

6: AML Pg. 14, Ex. 1.6

Answer:

a) Recommending a book to a user in an online bookstore is most suited to supervised learning. The input data would be the user's purchase history and the target output data would be whether or not the user bought the book. The model would try to predict the likelihood of the user buying the book, which would indicate whether the book should be recommended. Reinforcement learning is also suitable. Again, the input data would be the user's purchase history. The output action would be a recommendation. The grade for that action would be low/negative if the user didn't buy the suggested book, and high/positive if the user bought the suggested book.

b) Playing tic-tac-toe is most suited to reinforcement learning. The input data would be each move taken, the output would be the result of that move (possibly the follow-up action of the opposition), and the grade of that output would be how poorly off you are after the opposition makes a move following your move.

c) Categorizing movies into different types can be done either via supervised or unsupervised learning. With either, the input would be a list of characters, director, title, and more metadata. Or, if the algorithm were advanced enough, the input could be the individual frames of the movie itself. In the supervised case, the target output would be one of however many categories you choose as possible movie types. In the unsupervised case, the movies would be divided into categories that a human would then have to identify. These categories could be based on movie type, but may also be based on sets of actors or directors. One cannot know until the learning algorithm is run.

d) Learning to play music is most suited to reinforcement learning. The input would be information about the note to play: its intensity and distribution of frequencies. The output would be the actual production of a note. A grade would be assigned to this output according to how well it matched the desired intensity and distribution of frequencies.

e) Deciding the maximum allowed debt for each bank customer is most suited to supervised learning. The input would be a profile of the customer's financial history. The output would be a debt allowance, and the target output would be what the bank would actually give that customer. This task could also be done with reinforcement learning. The input and output would be the same as above. The grade assigned to the output would be an estimate of how much such a debt allowance would make or lose for the bank for the given customer.

7: AML Pg. 23, Ex. 1.10

Answer:

a) The μ for each of the three coins selected is 0.5, assuming each is a fair coin.

b) See code.

c) See code.

d) The first and the randomly selected coins obey the Hoeffding bound by chance – they are randomly-drawn size-10 samples from bins whose probabilities of heads are 0.5. The coin selected to have a minimum v does not obey the Hoeffding bound because, by chance and our selection, it is a sample whose probability of heads most drastically differs from the probability of heads of its bin (0.5). If the bin represents a hypothesis, this is analogous to the problem that the Hoeffding bound *should* linearly scale with the number of hypotheses, though as plotted it does not.

e) Each bin in Figure 1.10 here represents a coin (1,000 coins means 1,000 bins). Each marble is a coin flip. On average, then, there should be equal numbers of red and green marbles (heads and tails) in each bin. That is, for any bin i, $\mu_i = 0.5$. When we flip a coin for a sample 10 times, we would analogously be randomly drawing 10 marbles from the bin. The 10,000 experiment runs correspond to 10,000 times drawing the aforementioned samples for each of the 1,000 bins. There is no good analog here between error and coin flips. In-sample error sums the number of times the hypothesis function disagrees with the target function over the N sample inputs. The key metric here, however, is the difference between the *average* of N sample inputs (i.e. probability of heads) and the underlying probability of heads (0.5). Out-of-sample error is not a valid concept here, as there is no real hypothesis for any bin – just the probability of heads for a given sample, which we know is 0.5 for each bin's population.

8: PRML Pg. 58, Ex. 1.1

Answer: The weights that minimize E(W) are given by

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

$$= \sum_{n=1}^{N}[\sum_{j=0}^{M} w_j x_n^j - t_n]x_n^i = 0$$

$$\sum_{n=1}^{N}[\sum_{j=0}^{M} w_j x_n^i] = \sum_{n=1}^{N} x_n^i t_n$$

By equations 1.122, the RHS becomes as follows:

$$\sum_{n=1}^{N}[\sum_{j=0}^{M} w_j x_n^i x_n^i] = T_i$$

$$\sum_{n=1}^{N}[\sum_{j=0}^{M} w_j x_n^{i+j}] = T_i$$

$$\sum_{j=0}^{M} w_j \sum_{n=1}^{N} x_n^{i+j} = T_i$$

By equations 1.122, the LHS becomes as follows:

$$\sum_{j=0}^{M} w_j A_{ij} = T_i$$

The weights minimizing the error function are therefore given by the solutions to the system of linear equations of 1.122 and 1.123:

$$\sum_{j=0}^{M} A_{ij} w_j = T_i$$

9: PRML Pg. 58, Ex. 1.5

Answer: Equation 1.38:

$$\begin{aligned} var(f) &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)\mathbb{E}[f(x)]] + \mathbb{E}[\mathbb{E}[f(x)]^2] \end{aligned}$$

But the expectation value of a constant (here, another expectation value), is the constant itself.

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

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

So equation 1.38 satisfying equation 1.39, which is the following:

$$var(f) = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

10: PRML Pg. 59, Ex. 1.6

Answer: Covariance is defined as follows:

$$cov(x, y) = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

$\mathbb{E}_{x,y}[xy]$ is likewise defined as follows, where p(x) is x's probability density.

$$\mathbb{E}_{x,y}[xy] = \int_x \int_y p(xy)\, xy \, \mathrm{d}x \, \mathrm{d}y$$

If x and y are independent, then, by definition, $p(xy) = p(x)p(y)$. Therefore...

$$\mathbb{E}_{x,y}[xy] = \int_x p(x)\,x\,\mathrm{d}x \int_y p(y)\,y\,\mathrm{d}y$$

$$\mathbb{E}_{x,y}[xy] = \mathbb{E}[x]\mathbb{E}[y]$$

The covariance for two independent variables x and y is therefore 0:

$$cov(x, y) = \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] = 0$$

11: PRML Pg. 59, Ex. 1.9

Answer: The maximum of the Gaussian distribution is the x at which the derivative of $\mathcal{N}$ equals 0. That is,

$$\frac{\mathrm{d}}{\mathrm{dx}}\mathcal{N}(x\,|\,\mu,\sigma) = \frac{1}{2\pi\sigma^2}(2\frac{-1}{2\sigma^2}(x-\mu))e^{\frac{-1}{2\sigma^2}(x-\mu)^2} = 0$$

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

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

$$x = \mu$$

Likewise, the maximum of a multivariate Gaussian distribution is the $\boldsymbol{x}$ at which the derivative of $\mathcal{N}$ equals 0. That is,

$$\nabla_x\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})\nabla_x[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] = 0$$

$$= -\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = 0$$

$$\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

$$\boldsymbol{x} = \boldsymbol{\mu}$$

12: PRML Pg. 59, Ex. 1.11

Answer:

$$\ln p(\boldsymbol{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln 2\pi$$

Finding the log-likelihood-maximizing μ...

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\ln p(\boldsymbol{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}2(x_n-\mu)(-1) = 0$$

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

$$\mu = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n$$

Finding the log-likelihood-maximizing σ^2...

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\ln p(\boldsymbol{x}|\mu,\sigma^2) = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\frac{1}{\sigma^2} = 0$$

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

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

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

13: PRML Pg. 60, Ex. 1.12

Answer: If $n = m$, by equation 1.50...

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n^2] = \mu^2 + \sigma^2$$

If $n \neq m$, x_n and x_m are independently sampled, so

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n][x_m] = \mu\mu = \mu^2$$

From equation 1.55, we have $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$. This is a constant and the definition of the mean of a set of N data points. Since these x_n are distributed normally, the expectation value of this max-likelihood mean is also the mean of the Gaussian dataset.

$$\mathbb{E}[\mu_{ML}] = \mu$$

From equation 1.56, we have $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$

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

$$= \frac{1}{N}\sum_{n=1}^{N}(x_n^2 - 2\frac{x_n}{N}\sum_{m=1}^{N} x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N} x_m x_l)$$

Because the expectation value is the linear sum of expectation values, we can distribute the expectation into the sum over n terms:

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}(\mathbb{E}[x_n^2] - 2\mathbb{E}[\frac{x_n}{N}\sum_{m=1}^{N} x_m] + \mathbb{E}[\frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N} x_m x_l])$$

The first term is given by equation 1.50. The expectation of the second term is given by equation 1.130 (see explanation below). The expectation of the third term, as above, can be rewritten as the double sum of $\mathbb{E}[x_m x_l]$.

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}((\mu^2+\sigma^2) - 2(\mu^2+\frac{1}{N}\sigma^2) + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbb{E}[x_m x_l])$$

The third term is given by equation 1.130. The $\frac{1}{N}\sigma^2$ term appears because n=m N times in the course of the double summation, so $\frac{1}{N^2} * N * \sigma^2 = \frac{1}{N}\sigma^2$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}((\mu^2+\sigma^2) - 2(\mu^2+\frac{1}{N}\sigma^2) + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}(\mu^2) + \frac{1}{N}\sigma^2)$$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}((\mu^2+\sigma^2) - 2(\mu^2+\frac{1}{N}\sigma^2) + \frac{1}{N^2}\sum_{m=1}^{N}\mu^2(N+1-1) + \frac{1}{N}\sigma^2)$$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}((\mu^2+\sigma^2) - 2(\mu^2+\frac{1}{N}\sigma^2) + \frac{1}{N^2}\mu^2(N+1-1)(N+1-1) + \frac{1}{N}\sigma^2)$$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}((\mu^2+\sigma^2) - 2(\mu^2+\frac{1}{N}\sigma^2) + \mu^2 + \frac{1}{N}\sigma^2)$$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{1}{N}\sum_{n=1}^{N}(\sigma^2 + \frac{1}{N}\sigma^2 + \frac{1}{N}\sigma^2)$$

$$\mathbb{E}[\sigma^2_{ML}] = \frac{N-1}{N}\sigma^2$$