Problem Set 1

AML 1.3

Hint #1. For part (a), read the paragraph in AML that precedes Eq. 1.3 on pg. 7. Specifically, the algorithm picks an example that is currently misclassified. Note that $y(t)$ can take two values, so analyze the cases.

AML 1.10

Hint #1. Read AML Sec. 1.3.2, and review Lecture 2: Is Learning Feasible?.

PRML 1.5

Hint #1. The variance of $f(x)$ is defined by Eq. 1.38; expand the square.

Hint #2. Use the linearity property of the expectation operator: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

PRML 1.6

Hint #1. For two random variables x and y, the covariance is defined by Eq. 1.41.

Hint #2. Use Eq. 1.33 and the fact that $p(x, y) = p(x)p(y)$ when x and y are independent to rewrite the expression $\mathbb{E}[xy]$ (which is one of the expressions on the right hand side of Eq. 1.41) and show the result for the discrete case.

Hint #3. Rinse and repeat for the case where x and y are continuous.

PRML 1.11

Hint #1. Use standard rules of differentiation to take the derivative of Eq. 1.54 with respect to μ; set this expression equal to zero; and solve for μ.

Hint #2. Use standard rules of differentiation to take the derivative of Eq. 1.54 with respect to σ^2; set this expression equal to zero; and solve for σ^2.

Problem Set 2

AML 3.1

Hint #1. We answered this question in class. Your answer should explain *why*.

AML 3.7

Hint #1. Your starting point $E_{\text{in}}(\mathbf{w})$ is given by Eq. 3.9 on pg. 91.

AML 3.8

Hint #1. Pay close attention to the third paragraph in Sec. 3.3.2 on pg. 93. Your answer should involve the Taylor expansion.

AML 3.9

You can generate a dataset that is not linearly separable, train a perceptron, and compute the different pointwise error measures. But this problem can (and should) be solved analytically.

Hint #1. For part (a), the problem statement says $y = +1$. You should plot the three pointwise error measures on a plot of $\mathbf{e}$-vs-s, i.e., the x-axis is s (the value of the signal) and the y-axis is the value of the pointwise error. So, for example, if $y = +1$, then you can plot $\mathbf{e}_{\text{class}}$ for different values of s. Note that $\mathbf{e}_{\text{class}}$ is defined using the indicator function, so your plot of $\mathbf{e}_{\text{class}}$ is going to be a discontinuous graph.

Hint #2. For part (b), analyze the cases (similar to analysis we did in class when we examined the cross-entropy criterion) when (i) the signal and target agree, i.e., the product of y and s is greater than or equal to zero (what can you say about $\mathbf{e}_{\text{class}}$ and $\mathbf{e}_{\text{sq}}$?) and (ii) the signal and target disagree, i.e., the product of y and s is less than or equal to zero.

Hint #3. Part (c) is like part (b).

PRML 2.43

Hint #1. The problem statement indicates that Eq. 2.293 is a generalization of the univariate Gaussian distribution. Of course, the Gaussian distribution is symmetric. By inspection, Eq. 2.293 is centered at 0. Therefore, we can make the following change:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{|x|^q}{2\sigma^2}\right) dx = 2\int_{0}^{\infty} \exp\left(-\frac{x^q}{2\sigma^2}\right) dx \tag{1}$$

Hint #2. Use a change of variables by letting $u = \frac{x^q}{2\sigma^2}$. With this change of variables, we can solve for x as a function of u and compute $\frac{du}{dx}$.

Hint #3. The Gamma function is defined as follows:

$$\Gamma(x) \equiv \int_{0}^{\infty} u^{x-1} e^{-u} du \tag{2}$$

Hints 1–3 are very useful for proving Eq. 2.294 where $p(x|\sigma^2, q)$ is given by Eq. 2.293. To show that Eq.2.293 reduces to the Gaussian when $q = 2$, make the substitution and compare it to Eq. 2.42. The last part of the question asks you to show that the log likelihood function over $\mathbf{w}$ and σ^2, for a given $\mathbf{X}$ and $\mathbf{t}$, is given by Eq. 2.295. This problem broaches a very, very important pattern in machine learning, which is building and optimizing a likelihood function. Here is the technique:

1. **Define the likelihood function.** Eq. 1.53 is a prototypical example of how to define a likelihood function; the example uses a Gaussian distribution but the approach is the same for any distribution (like Eq. 2.293). You simply take your dataset $\mathbf{X}$ and use your probability distribution to compute the probability of each input vector $\mathbf{x}_n$, and the likelihood function is the product of these probabilities.

2. **Take the natural log of both sides to get the log likelihood function.** We take the natural log of both sides because products turn into sums (i.e., $\prod$ becomes $\sum$) and other important properties of the natural log break up the log likelihood function into a sum of terms, e.g., Eq. 2.295.

3. **Take the derivative of the log likelihood function.**

4. **Set the derivative equal to zero and solve for the parameters.**

PRML 3.2

Hint #1. There is a right way to think about a matrix-vector product: if $b = Ax$, then b is a linear combination of the columns of A. So, rather than the familiar take on matrix-vector products, which is $b_i = \sum_j a_{ij}x_j$, think of matrix-vector products as $b = Ax = \sum_j x_j a_j$, where a_j are the columns of A. In class, I drew this linear combination on the board.

Hint #2. Multiplying a matrix and a vector produces a vector. Multiplying this new vector by another matrix produces another vector. For example, suppose A and B are matrices and x, b, and c are vectors. Then, we may have $BAx \Rightarrow c = Bb$.

Matrix-vector products are fundamentally important in neural networks. If you are interested, a good source for foundational topics in numerical linear algebra such as matrix-vector products and norms is the Trefethen and Bau masterpiece *Numerical Linear Algebra.*

PRML 4.1

Hint #1. In class, I drew what a convex hull looks like for a set of points $\mathbf{x}_n$. Then I drew another convex hull for another set of points $\mathbf{y}_m$. Then I emphasized that if two datasets are linearly separable, then that means you can draw a separating hyperplane between the two sets of points, but it also means that no point lies in the intersection of the convex hulls. If a point z lies in the intersection of the convex hulls, then the point must satisfy the following:

$$z = \sum_n \alpha_n x_n = \sum_m \beta_m y_m \tag{3}$$

where $\beta_m \geq 0$ and $\sum_m \beta_m = 1$. So I said to show that if two convex hulls intersect (i.e., there exists a point z) then show the contradiction that a separating hyperplane cannot possibly have all the x_n on one side and all the y_m on the other.

PRML 4.14

Hint #1. First, review the four steps for building and optimizing a likelihood function enumerated above. Then notice that the likelihood function for the logistic regression model is given by Eq. 4.89. Then they take the natural log of both sides to get the log likelihood function, which breaks up the product into a sum. Negating this quantity turns into an error measure that we can optimize (specifically, minimize). So, we take the gradient of Eq. 4.90 with respect to $\mathbf{w}$ and get Eq. 4.91. The gradient is optimized when the y_n equal the t_n, right? When does this happen? Note that $y_n = \sigma(\mathbf{w}^T x_n)$, and we reviewed the logistic function in class. Incidentally, this problem has very important practical ramifications. Keep this problem in mind when we review regularization techniques.

PRML 4.18

Hint #1. Use Eq. 4.108 to compute $\frac{\partial E}{\partial y_{nk}}$. Then combine this simple expression with Eq. 4.106 using the chain rule. Recall the chain rule is something like $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial y}$.

Problem Set 3

AML 3.2

Hint #1. You already implemented the perceptron learning algorithm in AML Exercise 1.4. There are only a few trivial differences between what you implemented for that exercise and what you need to do for this exercise:

1. Generate a *non*-separable dataset. The specification for Ex. 3.2 tells you how to do this.
2. Per the PA on pg. 80 (Line 4), evaluate all examples using $\mathbf{w}(t+1)$ to get $E_{\text{in}}(\mathbf{w}(t+1))$.
3. Per the PA on pg. 80 (Line 5), cache the best weight vector encountered.

PRML 1.2

Hint #1. Almost exactly as you did for PRML 1.1 in a previous problem set, consider the error function given by Eq. 1.4 in which the function $y(x, \mathbf{w})$ is given by the polynomial Eq. 1.2:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - t_n\right)^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \tag{4}$$

where $||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$. As the problem statement suggests (and as you did for PRML 1.1), take the derivative with respect to w_i. The only difference from PRML 1.1 is you have an additional (simple) term to take the derivative of. Note that λ is a constant. So you should end up with something that looks like Eq. 1.122 except the form of A_{ij} is going to be slightly different.

PRML 1.3

Joint probabilities, conditional probabilities, and marginal probabilities are very important quantities in machine learning and neural networks in particular. The first part of Sec. 1.2 is an accessible introduction to the quantities. To solve PRML 1.3, suppose you have two random variables X and Y. The random variable X can take one of three values: $r$ed, $b$lue, or $g$reen. The random variable Y can take one of three values: $a$pple, $o$range, or $l$ime. There are two questions posed in this problem.

Hint #1. The first question is asking you to compute $p(Y = a)$. Use the sum rule (Eq. 1.10) and then the product rule (Eq. 1.11) to find an expression (a sum of three terms) for the marginal distribution $p(Y = a)$. You will be able to (easily) compute the components of each one of these terms using the counts and probabilities in the problem statement.

Hint #2. The second question is asking you to compute $p(X = g|Y = o)$. Use Bayes' theorem:

$$p(X = g|Y = o) = \frac{p(Y = o|X = g)p(X = g)}{p(Y = o)} \tag{5}$$

Computing the numerator is trivial. Computing the denominator is similar to the first question.

PRML 1.39

Hint #1. Use Tab. 1.3 to compute $p(x)$, $p(y)$, $p(x|y)$, and $p(y|x)$.

Hint #2. The entropy, conditional entropy, and mutual information are defined as follows:

$$H(x) = -\sum_i p(x_i) \ln p(x_i) \tag{6}$$

$$H(x|y) = -\sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j) \tag{7}$$

$$I(x, y) = H(x) - H(x|y) = H(y) - H(y|x) \tag{8}$$

PRML 2.12

Hint #1. Verifying the distribution is normalized is trivial.

Hint #2. To find the mean, note that the problem explicitly says *for a continuous variable*. How do you compute the expectation in the case of continuous variables? See Eq. 1.34.

Hint #3. To find the variance, note that you just computed $\mathbb{E}[x]$ and recall PRML 1.5, which you worked in a previous problem set. Compute the rest of what you need to get the variance using Eq. 1.40.

PRML 3.11

Hint #1. The uncertainty $\sigma_N^2(\mathbf{x})$ associated with the linear regression function is given by Eq. 3.59:

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}) \tag{9}$$

From Eq. (9), we have

$$\sigma_{N+1}^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_{N+1} \phi(\mathbf{x}) \tag{10}$$

where $\mathbf{S}_{N+1}$ is given by PRML 3.8:

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta \phi_{N+1} \phi_{N+1}^T \tag{11}$$

Now use Eq. (11), Eq. 3.110, and Eq. (9) to rewrite the r.h.s. of Eq. (10).

Hint #2. $\mathbf{S}_N$ is positive definite.

Problem Set 4

Note that PS4 problems are either WWW problems that only require you to fill in gaps or they are closely related to problems you already solved this semester.

PRML 5.1

Hint #1. The hint in the exercise says first find the relation between $\sigma(a)$ and $\tanh(a)$. You already did this in a previous problem set (see PRML 3.1). Then the hint says show that the parameters of the two networks differ by linear transformations.

PRML 5.7

Hint #1. You can take the derivative straightaway. This problem is similar to one you solved in a previous problem set (see PRML 4.17).

PRML 5.8

Hint #1. This exercise simply asks you to compute $\frac{\mathrm{d}}{\mathrm{d}a}\tanh(a)$, where $\tanh(a)$ is given by Eq. 5.59. Use the quotient rule and realize that your result should have the form given by Eq. 5.60. Thus, like the logistic sigmoid activation function, the derivative of the tanh activation function can be expressed in terms of the function value itself.

PRML 5.18

Hint #1. Suppose the extra parameters corresponding to skip-layer connections that go directly from the inputs to the outputs is given by the matrix s_{ij}. Introducing skip layer weights s_{ij} into the two-layer network of the form shown in Fig. 5.1 would only affect the forward propagation equation Eq. 5.64:

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} s_{ki} x_i \tag{12}$$

Now, write down the equations for the derivatives of the error function (Eq. 5.61) with respect to these additional parameters (s_{ki}).