1: AML Pg. 80, Ex. 3.2

Answer:

Note: When copying diagrams from Jupyter, axis labels of first two figures didn't render. Both have an x-axis of x1 value and a y-axis of x2 value.

In all the below graphs, yellow corresponds to a prediction/target value of 1, and purple corresponds to a prediction/target value of -1. Whether coloration denotes targets or predictions is specified in each figure's caption.

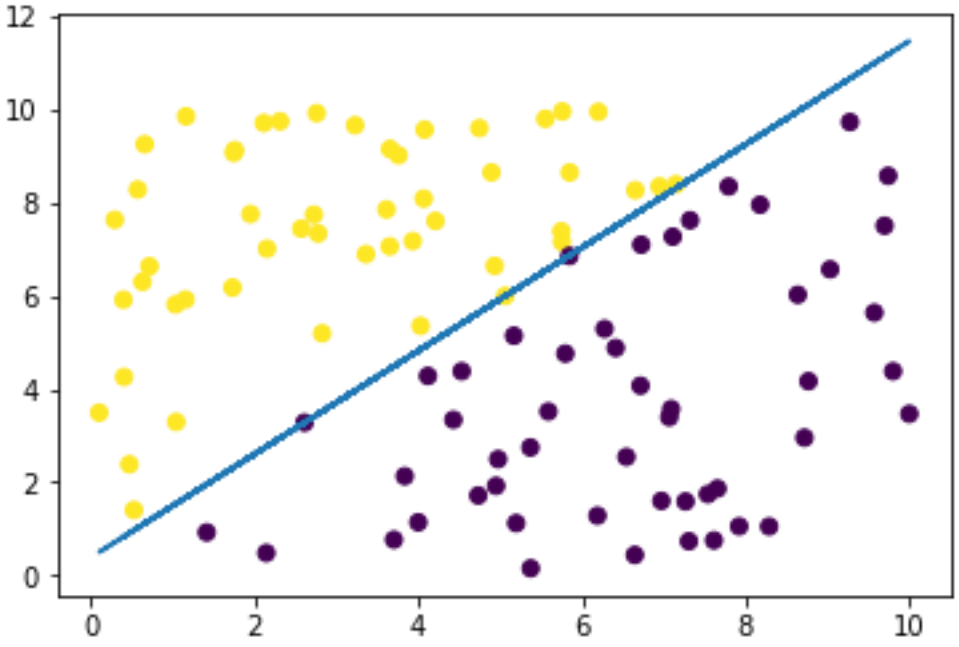

Figure 1: Target decision boundary and target data

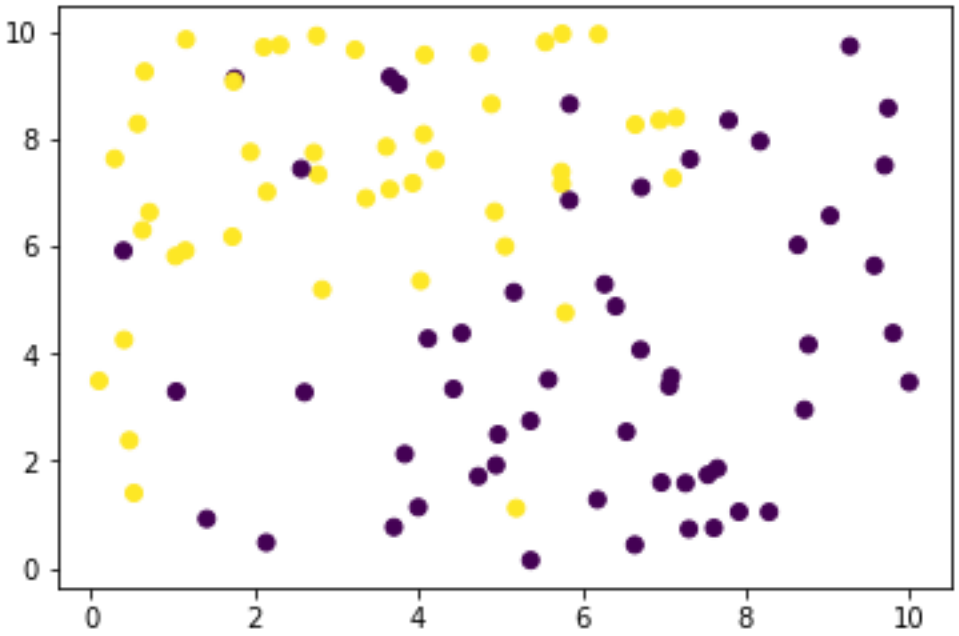

Figure 2: Target data after random negating of 10 targets

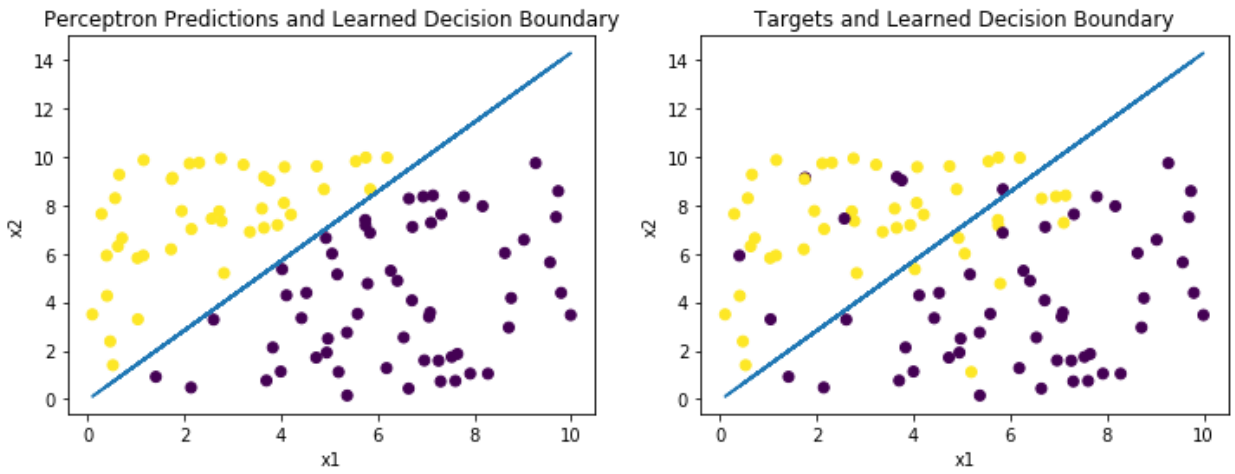

Figure 3: Decision boundary and perceptron predictions (left); Decision boundary and target data (right)

Figure 4 shows the average mean-squared errors (MSE) of perceptron predictions on the training set at each of 1000 iterations. The orange line shows the error of a perceptron whose weights are only updated in an iteration if the new weights would result in lower MSE. This explains why the orange line trends monotonically downward. Another explanation is, of course, that averaging over increasing time results in a downward pressure on the line.

The blue line, in contrast, shows the error of a perceptron whose weights are updated in each iteration, regardless of MSE with the new weights. Clearly, this line fluctuates between falling and dramatically rising MSE as the iterations go on. Were the number of iterations not set at

1000, this would go on forever, as the data are not linearly separable. Further, the overall increase in MSE arises because PLA is greedy – it is not guaranteed to improve overall MSE after a weight update. That is, an update in one iteration may bring the perceptron closer to correctly predicting a datapoint, but if that datapoint is an outlier (surrounded by targets of the opposite class), such an update may result in many more newly misclassified than correctly classified points. Note that an overall MSE increase as iterations increase is not guaranteed; rather, it depends on initial weights and the random selection of points on which to update.

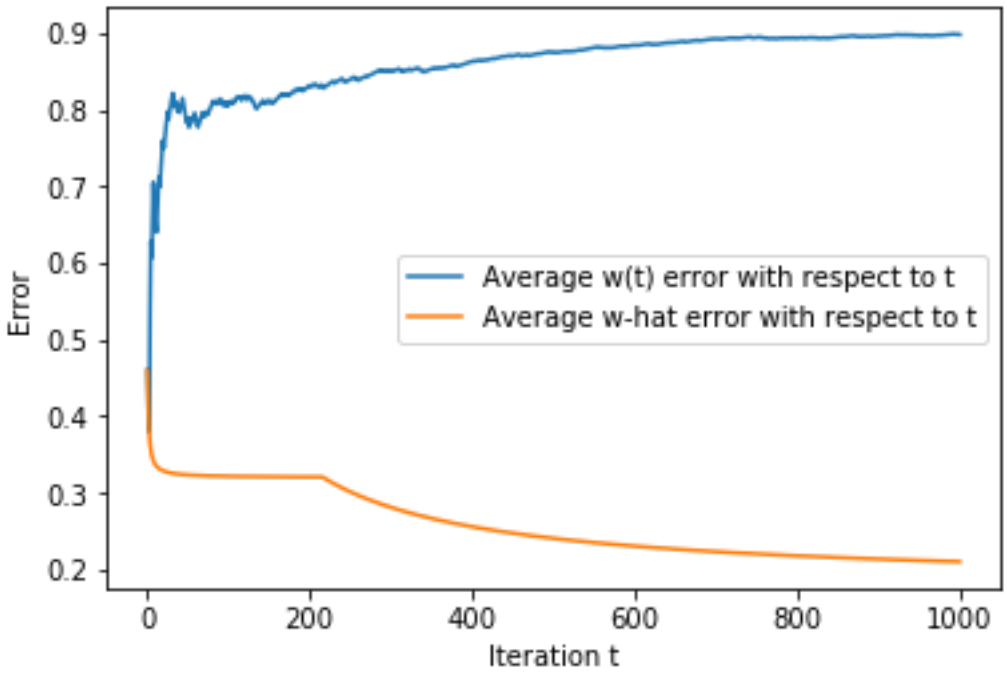

Figure 4: Average errors (over number of iterations) of pocket and regular PLA algorithms

Figure 5 shows the same phenomenon as described above, but with an overall decrease in MSE for the standard PLA (blue) line. Most likely, the MSE is actually stagnant, but more iterations means a decrease in time-averaged MSE.

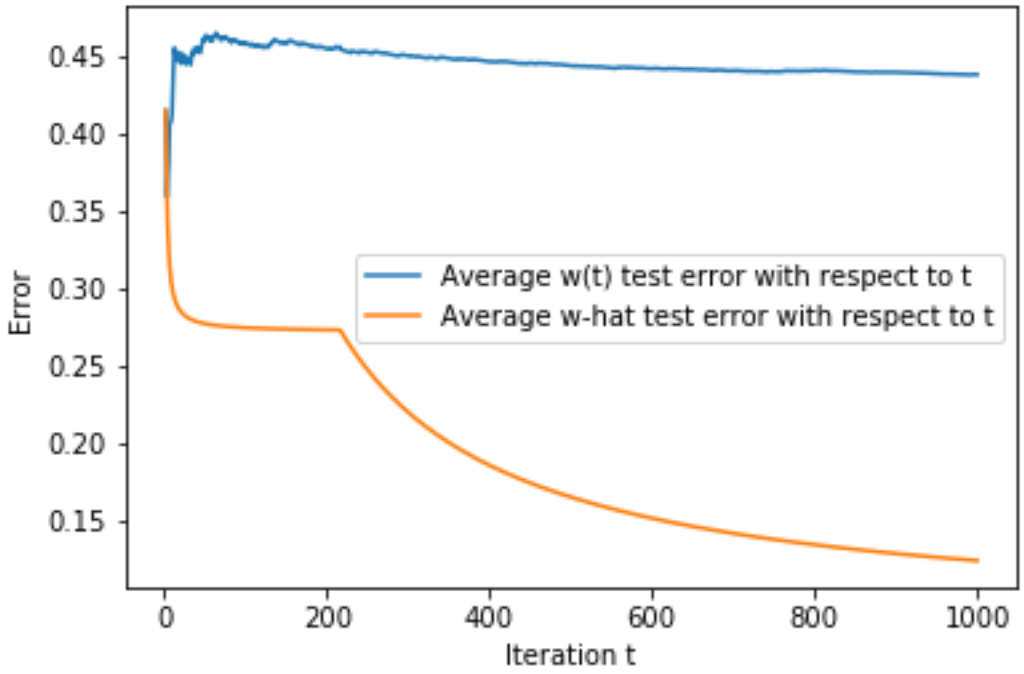

Figure 5: Average errors (over number of iterations) of pocket and regular PLA algorithms on the test set

2: AML Pg. 125, Ex. 4.3

Answer:

a) If the hypothesis set is fixed, but we increase the complexity of the target function, there will be more area between the best-fit curve and the target function. This is because the target function will fluctuate more with respect to the covariates, while the best fit will remain largely the same shape, limited by its degrees of freedom. The larger areas between the curves mean more deterministic noise.

For a fixed hypothesis set, an increase in target complexity leads to an increase in deterministic noise, which the model tries to fit, resulting in more overfitting.

b) If the target function is fixed and we decrease the complexity of the hypothesis set, we will increase deterministic noise. This is because the best fit will be worse of an approximation of the target (more area between their curves).

This increase in deterministic noise (i.e. an increase in bias) weighs towards higher expected out-of-sample error, but another affect dominates: decreasing hypothesis set complexity decreases variance (i.e. noise-fitting), lowering expected out-of-sample error overall. Lower expected out-of-sample error means less overfitting.

3: AML Pg. 142, Ex. 4.9

Answer:

Both curves increase with K because a larger validation set means a smaller training set. Specifically, a smaller training set increases out-of-sample and validation set error by increasing in-sample error. This is because of the VC Bound (Equation 4.1), where $\Omega(\mathcal{H})$ is constant for a given hypothesis set:

$$E_{out}(h) \leq E_{in}(h) + \Omega(\mathcal{H})$$

The two curves converge because a larger K increases the reliability of $E_{val}(g_{m^*}^-)$ as an estimate of $E_{out}(g_{m^*}^-)$. Mathematically, increasing K reduces the O term in the VC Bound (Eq. 4.11), decreasing the difference between $E_{val}(g_{m^*}^-)$ and $E_{out}(g_{m^*}^-)$:

$$E_{out}(g_{m^*}^-) \leq E_{val}(g_{m^*}^-) + O(\sqrt{\frac{\ln M}{K}})$$

4: PRML Pg. 58, Ex. 1.2

Answer:

$$E_{reg}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

where $y(x_n, \mathbf{w})$ is the prediction, t_n is the target, $\mathbf{w}$ is a vector of M weights, and there are N examples (data in the training set), each labeled n. Equivalently, the error can be written as follows:

$$E_{reg}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}[(\sum_{m=1}^{M} w_m x_n^{(m)} - t_n)^2] + \sum_{m=1}^{M} w_m^2$$

where each m denotes a covariate.

To minimize this regularized error function, we take its partial derivative with respect to an arbitrary weight w_i and set the derivative equal to 0:

$$\frac{\partial}{\partial w_i}E_{reg}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}[\frac{\partial}{\partial w_i}(\sum_{m=0}^{M} w_m x_n^{(m)} - t_n)^2] + \frac{\partial}{\partial w_i}\sum_{m=0}^{M} w_m^2 = 0$$

where we have moved the derivative into the summation over n because the weights do not vary with training examples. Applying the chain rule and recognizing that $\frac{\partial}{\partial w_i}\sum_{m=0}^{M} w_m^2 = 2\sum_{m=1}^{M} w_m$, we get the following:

$$\frac{\partial}{\partial w_i}E_{reg}(\mathbf{w}) = \sum_{n=1}^{N}[(\sum_{m=0}^{M} w_m x_n^{(m)} - t_n)x_n^{(i)}] + 2\sum_{m=0}^{M} w_m = 0$$

Distributing in $x_n^{(i)}$ and separating the sums...

$$\frac{\partial}{\partial w_i}E_{reg}(\mathbf{w}) = \sum_{n=1}^{N}\sum_{m=0}^{M} w_m x_n^{(m+i)} - \sum_{n=1}^{N} t_n x_n^{(i)} + 2\sum_{m=0}^{M} w_m = 0$$

Let us define the following terms:

$$A_{im} = \sum_{n=1}^{N} x_n^{(m+i)}$$

and

$$T_i = \sum_{n=1}^{N} t_n x_n^{(i)} - 2 \sum_{m=0}^{M} w_m$$

Then we can add T_i to both sides of the derivative to find the following:

$$\sum_{m=0}^{M} A_{im} w_m = T_i$$

5: PRML Pg. 58, Ex. 1.3

Answer:

Box r: 3 apples, 4 oranges, 3 limes
Box b: 1 apple, 1 orange, 0 limes
Box g: 3 apples, 3 oranges, 4 limes

$p(r) = 0.2$
$p(b) = 0.2$
$p(g) = 0.6$

$$p(apple) = p(r) * p(apple|r) + p(b) * p(apple|b) + p(g) * p(apple|g)$$
$$p(apple) = 0.2 * \frac{3}{10} + 0.2 * \frac{1}{2} + 0.6 * \frac{3}{10}$$
$$p(apple) = 0.34 = 34\%$$

$$p(g|orange) = \frac{p(orange|g) * p(g)}{p(orange)}$$
$$p(g|orange) = \frac{p(orange|g) * p(g)}{p(r) * p(orange|r) + p(b) * p(orange|b) + p(g) * p(orange|g)}$$
$$p(g|orange) = \frac{\frac{3}{10} * 0.6}{0.2 * \frac{4}{10} + 0.2 * \frac{1}{2} + 0.6 * \frac{3}{10}}$$
$$p(g|orange) = \frac{0.18}{0.36}$$
$$p(g|orange) = 0.5 = 50\%$$

6: PRML Pg. 65, Ex. 1.39

Answer:

First, calculate the probabilities. Note that $p(y|x) = \sum_i p(y, x_i)$

$$p(x=0) = p(x=0, y=0) + p(x=0, y=1)$$
$$p(x=0) = \frac{1}{3} + \frac{1}{3}$$
$$p(x=0) = \frac{2}{3} \tag{1}$$

$$p(x=1) = p(x=1, y=0) + p(x=1, y=1)$$
$$p(x=1) = 0 + \frac{1}{3} = \frac{1}{3}$$
$$p(x=1) = \frac{1}{3} \tag{2}$$

$$p(y=0) = \frac{1}{3} + 0$$
$$p(y=0) = \frac{1}{3} \tag{3}$$

$$p(y=1) = \frac{1}{3} + \frac{1}{3}$$
$$p(y=1) = \frac{2}{3} \tag{4}$$

Now for the conditional probabilities. Note that $p(y|x) = \frac{p(y,x)}{p(x)}$.

$$p(x=0|y=0) = \frac{p(x=0, y=0)}{p(y=0)}$$
$$p(x=0|y=0) = \frac{\frac{1}{3}}{\frac{1}{3}}$$
$$p(x=0|y=0) = 1 \tag{5}$$

$$p(x=0|y=1) = \frac{p(x=0, y=1)}{p(y=1)}$$
$$p(x=0|y=1) = \frac{\frac{1}{3}}{\frac{2}{3}}$$
$$p(x=0|y=1) = 0.5 \tag{6}$$

$$p(x=1|y=0) = \frac{0}{\frac{1}{3}}$$
$$p(x=1|y=0) = 0 \tag{7}$$

$$p(x=1|y=1) = \frac{\frac{1}{3}}{\frac{2}{3}}$$
$$p(x=1|y=1) = 0.5 \tag{8}$$

$$p(y=0|x=0) = \frac{\frac{1}{3}}{\frac{2}{3}}$$
$$p(y=0|x=0) = 0.5 \tag{9}$$

$$p(y=0|x=1) = \frac{0}{\frac{1}{3}}$$
$$p(y=0|x=1) = 0 \tag{10}$$

$$p(y=1|x=0) = \frac{\frac{1}{3}}{\frac{2}{3}}$$
$$p(y=1|x=0) = 0.5 \tag{11}$$

$$p(y=1|x=1) = \frac{\frac{1}{3}}{\frac{1}{3}}$$
$$p(y=1|x=1) = 1 \tag{12}$$

Entropy is defined as $\mathrm{H}[x] = -\sum_i p(x_i) \ln p(x_i)$

a) $\mathrm{H}[x] = -p(x=0) \ln p(x=0) - p(x=1) \ln p(x=1)$

$$\mathrm{H}[x] = -\frac{2}{3} \ln \frac{2}{3} - \frac{1}{3} \ln \frac{1}{3}$$
$$\mathrm{H}[x] = 0.2703 + 0.3662 = 0.6365$$

b) $\mathrm{H}[y] = -p(y=0) \ln p(y=0) - p(y=1) \ln p(y=1)$

$$\mathrm{H}[y] = -\frac{1}{3} \ln \frac{1}{3} - \frac{2}{3} \ln \frac{2}{3}$$
$$\mathrm{H}[y] = 0.3662 + 0.2703 = 0.6365$$

Conditional Entropy is defined as $\mathrm{H}[x|y] = -\sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j)$

c) $\mathrm{H}[y|x] = -\sum_i p(y_i, x=0) \ln p(y_i|x=0) + p(y_i, x=1) \ln p(y_i|x=1)$

$$\mathrm{H}[y|x] = -(p(y=0, x=0) \ln p(y=0|x=0) + p(y=0, x=1) \ln p(y=0|x=1) +$$
$$p(y=1, x=0) \ln p(y=1|x=0) + p(y=1, x=1) \ln p(y=1|x=1))$$
$$\mathrm{H}[y|x] = -(\frac{1}{3} \ln 0.5 + 0 \ln 0 + \frac{1}{3} \ln 0.5 + \frac{1}{3} \ln 1)$$
$$\mathrm{H}[y|x] = -(-0.2310 + 0 - 0.2310 + 0) = 0.4621$$

d) $\mathrm{H}[x|y] = -\sum_i p(x_i, y=0)\ln p(x_i|y=0) + p(x_i, y=1)\ln p(x_i|y=1)$

$$\mathrm{H}[x|y] = -(p(x=0,y=0)\ln p(x=0|y=0) + p(x=0,y=1)\ln p(x=0|y=1)$$

$$+p(x=1,y=0)\ln p(x=1|y=0) + p(x=1,y=1)\ln p(x=1|y=1))$$

$$\mathrm{H}[x|y] = -(\frac{1}{3}\ln 1 + \frac{1}{3}\ln 0.5 + 0\ln 0 + \frac{1}{3}\ln 0.5)$$

$$\mathrm{H}[x|y] = -(0 - 0.2310 + 0 - 0.2310) = 0.4621$$

Joint Entropy is defined as $\mathrm{H}[x,y] = -\sum_i \sum_j p(x_i, y_j)\ln p(x_i, y_j)$

e) $\mathrm{H}[x,y] = -\sum_i p(x_i, y=0)\ln p(x_i, y=0) + p(x_i, y=1)\ln p(x_i, y=1)$

$$\mathrm{H}[x,y] = -(p(x=0,y=0)\ln p(x=0,y=0) + p(x=0,y=1)\ln p(x=0,y=1) +$$

$$p(x=1,y=0)\ln p(x=1,y=0) + p(x=1,y=1)\ln p(x=1,y=1))$$

$$\mathrm{H}[x,y] = -(\frac{1}{3}\ln\frac{1}{3} + \frac{1}{3}\ln\frac{1}{3} + 0\ln 0 + \frac{1}{3}\ln\frac{1}{3})$$

$$\mathrm{H}[x,y] = -(3 * \frac{1}{3}\ln\frac{1}{3})$$

$$\mathrm{H}[x,y] = -\ln\frac{1}{3}$$

$$\mathrm{H}[x,y] = 1.0986$$

Mutual Information is defined as $\mathrm{I}[x,y] = \mathrm{H}[x] - \mathrm{H}[x|y] = \mathrm{H}[y] - \mathrm{H}[y|x]$

f) $\mathrm{I}[x,y] = 0.6365 - 0.4621 = 0.1744$

7: PRML Pg. 130, Ex. 2.12

Answer:

The uniform distribution

$$U(x|a,b) = \frac{1}{b-a} \qquad a \leq x \leq b$$

is normalized if

$$\int_a^b \frac{dx}{b-a} = 1$$

since x can only have values between a and b. Performing this integration, this is indeed the case:

$$\int_a^b \frac{dx}{b-a} = [\frac{x}{b-a}]_a^b$$

$$= \frac{b}{b-a} - \frac{a}{b-a}$$

$$= \frac{b-a}{b-a} = 1$$

The mean of x distributed via $U(x|a,b) = \frac{1}{b-a}$ is found by integrating the product of x and its distribution:

$$\begin{aligned}
\text{mean} = \mathbb{E}(x) &= \int_a^b x \frac{1}{b-a} dx \\
&= [\frac{x^2}{2(b-a)}]_a^b \\
&= \frac{b^2}{2(b-a)} - \frac{a^2}{2(b-a)} \\
&= \frac{b^2 - a^2}{2(b-a)} \\
&= \frac{(b-a)(b+a)}{2(b-a)} \\
&= \frac{b+a}{2}
\end{aligned}$$

The variance of x distributed via $U(x|a,b) = \frac{1}{b-a}$ is found with the mean derived above, giving the following:

$$\begin{aligned}
\text{var} &= \mathbb{E}(x^2) - \mathbb{E}(x)^2 \\
&= \int_a^b x^2 \frac{1}{b-a} dx - (\frac{b+a}{2})^2 \\
&= [\frac{x^3}{3(b-a)}]_a^b - (\frac{b+a}{2})^2 \\
&= [\frac{x^3}{3(b-a)}]_a^b - (\frac{b+a}{2})^2 \\
&= \frac{b^3 - a^3}{3(b-a)} - (\frac{b+a}{2})^2 \\
&= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} - (\frac{b+a}{2})^2 \\
&= \frac{(a^2+ab+b^2)}{3} - (\frac{b+a}{2})^2 \\
&= \frac{(a^2+ab+b^2)}{3} - \frac{(a^2+2ab+b^2)}{4} \\
&= \frac{\frac{4}{3}(a^2+ab+b^2)}{4} - \frac{(a^2+2ab+b^2)}{4} \\
&= \frac{\frac{1}{3}a^2 - \frac{2}{3}ab + \frac{1}{3}b^2}{4} \\
&= \frac{a^2 - 2ab + b^2}{12} \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

8: PRML Pg. 174, Ex. 3.4

Answer:

Sum of squares error:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(w_0 + \sum_{i=1}^{D} w_i x_n^{(i)} - t_n)^2$$

With a noise term $\epsilon_i \sim \mathcal{N}(0|\sigma^2)$ added independently to each input, this error becomes the following:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(w_0 + \sum_{i=1}^{D} w_i (x_n^{(i)} + \epsilon_i) - t_n)^2$$

To simplify things, define $g_n = w_0 + \sum_{i=1}^{D} w_i(x_n^{(i)} + \epsilon_i) = y_n + \sum_{i=1}^{D} w_i\epsilon_i$ where $y_n = w_0 + \sum_{i=1}^{D} w_i x_n^{(i)}$ is the prediction on noiseless data.

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(g_n - t_n)^2$$

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(g_n^2 - 2g_n t_n + t_n^2)$$

Substituting our g_n back in...

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}([y_n + \sum_{i=1}^{D} w_i\epsilon_i]^2 - 2[y_n + \sum_{i=1}^{D} w_i\epsilon_i]t_n + t_n^2)$$

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + 2y_n\sum_{i=1}^{D} w_i\epsilon_i + (\sum_{i=1}^{D} w_i\epsilon_i)^2 - 2y_n t_n - 2t_n\sum_{i=1}^{D} w_i\epsilon_i + t_n^2)$$

Taking the expectation of the error under the noise distribution, we find the following:

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}(y_n^2 + 2y_n\sum_{i=1}^{D} w_i\epsilon_i + (\sum_{i=1}^{D} w_i\epsilon_i)^2 - 2y_n t_n - 2t_n\sum_{i=1}^{D} w_i\epsilon_i + t_n^2)$$

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + 2y_n\mathbb{E}[\sum_{i=1}^{D} w_i\epsilon_i] + \mathbb{E}[(\sum_{i=1}^{D} w_i\epsilon_i)^2] - 2y_n t_n - 2t_n\mathbb{E}[\sum_{i=1}^{D} w_i\epsilon_i] + t_n^2)$$

We can move the expectations past the weights in the second and fifth terms because the weights are constant with respect to the noise distribution:

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + 2y_n\sum_{i=1}^{D} w_i\mathbb{E}[\epsilon_i] + \mathbb{E}[(\sum_{i=1}^{D} w_i\epsilon_i)^2] - 2y_n t_n - 2t_n\sum_{i=1}^{D} w_i\mathbb{E}[\epsilon_i] + t_n^2)$$

Because $\mathbb{E}[\epsilon_i] = 0$, the second and fifth terms are zero. Now to deal with the last expectation term...

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + \mathbb{E}[(\sum_{i=1}^{D} w_i\epsilon_i)^2] - 2y_n t_n + t_n^2)$$

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + \mathbb{E}[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \epsilon_i \epsilon_i] - 2y_n t_n + t_n^2)$$

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \mathbb{E}[\epsilon_i \epsilon_i] - 2y_n t_n + t_n^2)$$

Since we are given $\mathbb{E}[\epsilon_i\epsilon_j] = \delta_{ij}\sigma^2$, we find the following:

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + \sigma^2\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \delta_{ij} - 2y_n t_n + t_n^2)$$

This delta function indicates that the second term sum equals 0 for all $i \neq j$. It therefore can be rewritten as one sum over i:

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 + \sigma^2\sum_{i=1}^{D} w_i^2 - 2y_n t_n + t_n^2)$$

The second term is independent of the outer sum, so we can rewrite this as follows:

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n^2 - 2y_n t_n + t_n^2) + \frac{1}{2}\sigma^2\sum_{i=1}^{D} w_i^2$$

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2 + \frac{1}{2}\sigma^2\sum_{i=1}^{D} w_i^2$$

As shown above, the expectation over the noise distribution of the new error gives the old error $\frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2$ plus a regularization term $\frac{1}{2}\sigma^2\sum_{i=1}^{D} w_i^2$. That is, on average, minimizing the new error (noisy inputs) is equivalent to minimizing the old error but with a weight-decay term.

9: PRML Pg. 175, Ex. 3.11

Answer:

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x}) \tag{13}$$

From the web solutions to exercise 3.8, we have the following:

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\phi(x_{N+1})\phi(x_{N+1})^T$$

$$\mathbf{S}_{N+1} = [\mathbf{S}_N^{-1} + \beta\phi(x_{N+1})\phi(x_{N+1})^T]^{-1}$$
$$\mathbf{S}_{N+1} = [\mathbf{S}_N^{-1} + \sqrt{\beta}\phi(x_{N+1})\sqrt{\beta}\phi(x_{N+1})^T]^{-1}$$

Now equating the left-hand side of PRML identity 3.110 with the right-hand side of the above equation, we find the following:

$$\mathbf{S}_{N+1} = [\mathbf{S}_N^{-1} + \sqrt{\beta}\phi(x_{N+1})\sqrt{\beta}\phi(x_{N+1})^T]^{-1} = \mathbf{S}_N - \frac{(\mathbf{S}_N\sqrt{\beta}\phi(x_{N+1}))(\sqrt{\beta}\phi(x_{N+1})^T\mathbf{S}_N)}{1 + \sqrt{\beta}\phi(x_{N+1})^T\mathbf{S}_N\sqrt{\beta}\phi(x_{N+1})}$$

Simplifying the RHS fraction...

$$\mathbf{S}_{N+1} = \mathbf{S}_N - \frac{\beta\mathbf{S}_N\phi(x_{N+1})\phi(x_{N+1})^T\mathbf{S}_N}{1 + \beta\phi(x_{N+1})^T\mathbf{S}_N\phi(x_{N+1})} \tag{14}$$

We want to prove the following:

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x})$$

Or, equivalently:

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) \geq 0$$

To do so, use equation 13 (number refers to above equation, not PRML):

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

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

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \phi(\mathbf{x})^T(\mathbf{S}_N - \mathbf{S}_{N+1})\phi(\mathbf{x})$$

Now, substitute in equation 14 (number refers to above equation, not PRML):

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \phi(\mathbf{x})^T(\mathbf{S}_N - (\mathbf{S}_N - \frac{\beta\mathbf{S}_N\phi(x_{N+1})\phi(x_{N+1})^T\mathbf{S}_N}{1 + \beta\phi(x_{N+1})^T\mathbf{S}_N\phi(x_{N+1})}))\phi(\mathbf{x})$$

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \phi(\mathbf{x})^T(\frac{\beta\mathbf{S}_N\phi(x_{N+1})\phi(x_{N+1})^T\mathbf{S}_N}{1 + \beta\phi(x_{N+1})^T\mathbf{S}_N\phi(x_{N+1})})\phi(\mathbf{x})$$

Multiplying the numerator and denominator by $\frac{1}{\beta}$...

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \frac{\phi(\mathbf{x})^T\mathbf{S}_N\phi(x_{N+1})\phi(x_{N+1})^T\mathbf{S}_N\phi(\mathbf{x})}{\frac{1}{\beta} + \phi(x_{N+1})^T\mathbf{S}_N\phi(x_{N+1})}$$

Recognizing that the numerator is actually the square of three multiplied matrices, we can rewrite it as follows:

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \frac{[\phi(\mathbf{x})^T\mathbf{S}_N\phi(x_{N+1})]^2}{\frac{1}{\beta} + \phi(x_{N+1})^T\mathbf{S}_N\phi(x_{N+1})}$$

By properties of squaring, the numerator will always be positive. The multiplication $\phi(x_{N+1})^T\mathbf{S}_N\phi(x_{N+1})$ in the denominator must be positive in order for $\sigma_{N+1}^2(\mathbf{x})$ to be positive (Eq. 13), which is necssary by the properties of squaring and variance. Also, β is inverse nosie variance and therefore positive. The denominator is thus also positive, meaning $\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) \geq 0$ is true.