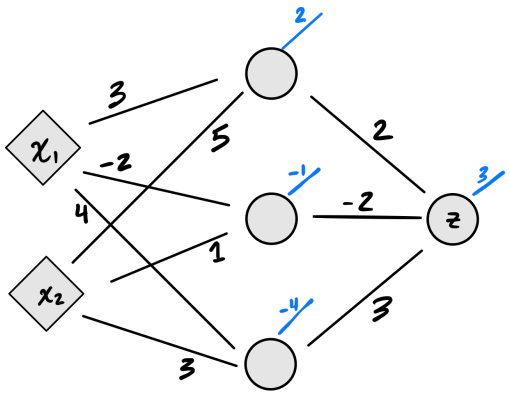
DSC 140A - Midterm 02 Review

Problem 1.

Consider the neural network shown below. The bias weight of the hidden neurons and the output neuron are shown as the blue numbers.

Given input $\vec{x} = (3, -1)^T$, what is the network's output? All activations are linear.



Solution: 46

Problem 2.

Let $H(\vec{x})$ be the neural network shown above. Suppose H is plotted. What will its plot look like?

- ◯ A straight line
- ◯ A curved line
- ● A plane in three dimensions
- ◯ A curved surface
- ◯ It is too high-dimensional to visualize

Solution: Because there are two input neurons and one output, $H : \mathbb{R}^2 \to \mathbb{R}$ Because all activations are linear, H is a linear function of its inputs. These two facts combined mean that the plot of H is a plane in three dimensions.

Problem 3.

Suppose the neural network above is modified so that the output neuron uses a sigmoid activation, but the hidden neurons still use linear activations. The output of the network will now be a number between 0 and 1.

Suppose the decision boundary is defined to be where the network outputs 1/2. True or False: the decision boundary can be non-linear.

Solution:

False.

The output neuron is using a non-linear activation, but that doesn't mean the decision boundary is automatically non-linear.

Consider the same exact network, except where the output neuron has a linear activation instead of a sigmoid activation. Call this network $J(\vec{x})$. Note that all activations in this network are linear, and so $J(\vec{x})$ is a linear function. If we plotted $J(\vec{x})$, it would be a plane.

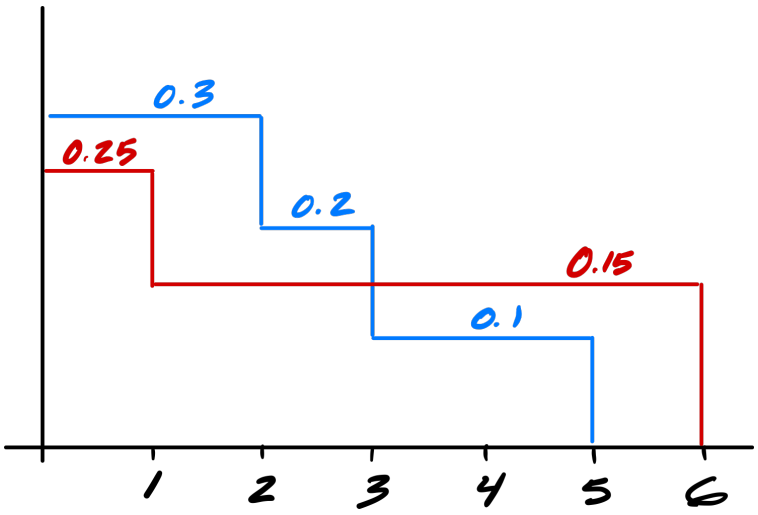
Recognize that $H(\vec{x})$ is just $\sigma(J(\vec{x}))$, where σ is the sigmoid function; that is, we take the output of the linear network J and feed it through the sigmoid function.

What does the decision boundary of H look like? This is the set of points $\vec{x}$ where $H(\vec{x}) = \sigma(J(\vec{x})) = 1/2$. Recall that the sigmoid function is $\sigma(z) = 1/(1 + e^{-z})$. Importantly, the output of the sigmoid is $1/2$ when the input is zero. As such, the decision boundary is where $J(\vec{x}) = 0$.

But J is a linear function, and so the set of $\vec{x}$ where $J(\vec{x}) = 0$ is just a straight line; the decision boundary is linear.

Problem 4.

Plotted below are two conditional densities, $p_1(x \mid Y = 1)$ and $p_0(x \mid Y = 0)$, describing the distribution of a continuous random feature for two classes, $Y = 0$ (red) and $Y = 1$ (blue).



Suppose $\mathbb{P}(Y = 1) = .4$ and $\mathbb{P}(Y = 0) = 0.6$.

a) What is the prediction of the Bayes classifier at $x = 2.5$?

Solution: Class 0 (red).

The prediction of the Bayes classifier is whichever class is larger between: $p_1(2.5 \mid Y = 1)\mathbb{P}(Y = 1)$ and $p_0(2.5 \mid Y = 0)\mathbb{P}(Y = 0)$.

We compute and compare:

$$\begin{aligned} p_1(2.5 \mid Y = 1)\mathbb{P}(Y = 1) &= (0.2)(0.4) \\ &= 0.08 \end{aligned}$$

$$p_0(2.5 \mid Y = 0)\mathbb{P}(Y = 0) = (0.15)(0.6)$$
$$= 0.09$$

Since Class 0 (red) is larger, we predict for Class 0.

b) What is the Bayes error for this distribution?

Solution: 0.37 (that is, 37%).

First, we need to find where the Bayes classifier predicts for Class 0, and where it predicts for Class 1.

We have already seen above that at $x = 2.5$, the Bayes classifier predicts Class 0 (red). It follows that it will also predict Class 0 everywhere between 2 and 3. In fact, it will also predict red for x between 3 and 6 as well. You can check that it will predict Class 0 between 0 and 1.

What about between 1 and 2? For example, what does the Bayes classifier predict at 1.5? We compute and compare:

$$p_1(1.5 \mid Y = 1)\mathbb{P}(Y = 1) = (0.3)(0.4)$$
$$= 0.12$$

$$p_0(1.5 \mid Y = 0)\mathbb{P}(Y = 0) = (0.15)(0.6)$$
$$= 0.09$$

And so here the prediction is for Class 1 (blue).

To summarize, the Bayes classifier predicts Class 1 (blue) between 1 and 2, and Class 0 (red) everywhere else.

The Bayes error is the probability of the classifier making a mistake. This can be broken into the probability of predicting Class 0 (red) when the correct answer was Class 0 (blue), plus the probability of predicting Class 1 (blue) when the correct answer was Class 0 (red).

In the first case, predicting class 0 when the right answer was 1, we have:

$$\mathbb{P}(\text{prediction} = 0 \text{ and } \mathrm{Y} = 1) = \mathbb{P}(\text{prediction} = 0 \mid \mathrm{Y} = 1)\mathbb{P}(Y = 1)$$

The probability that the prediction is 0 given that the point was drawn from the distribution for class 1 is the area under the blue curve in regions where the prediction is red (that is, between 0 and 1, and between 2 and 6). This area is:

$$0.3 + 0.2 + (0.1)(2) = 0.7.$$

And so:

$$\mathbb{P}(\text{prediction} = 0 \text{ and } \mathrm{Y} = 1) = \mathbb{P}(\text{prediction} = 0 \mid \mathrm{Y} = 1)\mathbb{P}(Y = 1)$$
$$= 0.7 \times 0.4$$
$$= 0.28$$

For the other case, where the prediction is 1 but $Y = 0$, we have:

$$\mathbb{P}(\text{prediction} = 1 \text{ and } \mathrm{Y} = 0) = \mathbb{P}(\text{prediction} = 1 \mid \mathrm{Y} = 0)\mathbb{P}(Y = 0)$$
$$= 0.15 \times 0.6$$
$$= 0.09,$$

where we got 0.15 from the area under the red curve where the Bayes classifier predicts blue; this is between 1 and 2.

In total, then, the Bayes error is $0.28 + 0.09 = 0.37$.

Problem 5.

Suppose a data set of n data points $\{(\vec{x}^{(i)}, y_i)\}$ are drawn from a probability distribution, and that the Bayes error on this distribution is 0.21. The data set is split into a training set of n_1 points and a test set of n_2 points.

Suppose a support vector machine (SVM) is trained on the training set and tested on the test set. True or False: it is possible for the SVM to obtain an 85% accuracy on the test set.

Solution: True.

The error rate of the Bayes classifier is a statement about the *probability* of a misclassification. That is: draw a new point from the distribution and use the Bayes classifier to predict its label. There is a 21% chance that the prediction is incorrect.

The Bayes error is *not* a guarantee about the performance of any classifier on a finite set of data points drawn from the distribution – maybe we got lucky or unlucky with our data and achieve better or worse than 79% accuracy.

What is true is that if we randomly sample a new point from the distribution, any classifier (e.g., and SVM) has *at least* a 21% probability of making a mistake. Or, to put it another way, if we generated infinitely many data sets of size n from the distribution, the average error rate of an SVM cannot be smaller than the Bayes error.

Problem 6.

Suppose a histogram is constructed for a set of n data points $\{x^{(i)}\}$. The histogram contains a bin $[2, 5)$ and the height of the histogram within this bin is 0.25.

Suppose a point is drawn uniformly at random from the data set. What is the probability that the point is in the interval $[2, 5)$?

Solution: 0.75, because that is the area under the histogram in $[2, 5)$.

Problem 7.

True or False. The number of negative entries in a covariance matrix must be even.

Solution: True. Diagonal entries cannot be negative, because they are variances. If an off-diagonal entry C_{ij} is negative, then C_{ji} must also be negative because covariance matrices are symmetric. This means that negative entries come in pairs, so there must be an even number of them. Note that zero is an even number, so this is true even when there are no negative entries.

Problem 8.

A data set contains 10 measurements of 50 different tumors. The covariance matrix containing the covariance between the features is computed. What is its shape?

Solution: 10×10

Problem 9.

The procedure below is used to make a classification between two classes. What is its name?

The data from the two classes are separated and full covariance matrices for each class are computed. These separate, full covariance matrices are combined into a single covariance matrix. The class means and the shared covariance matrix are used to estimate $p(x|Y = y)$ in a Bayes classifier.

- ● LDA
- ○ QDA
- ○ Naïve Bayes
- ○ Ridge Regression

Solution: It is LDA because the covariance matrices are combined to make a single shared covariance.

Problem 10.

The procedure below is used to make a classification. What is the procedure's name?

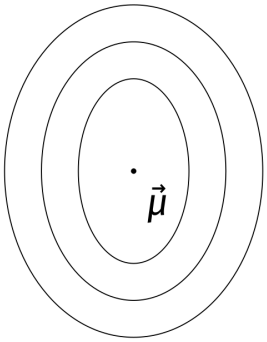
The data from two classes are separated and full covariance matrices for each class are computed independently (i.e., the covariance matrices are different). The class means and separate covariance matrices are used to estimate $p(x|Y = y)$ in a Bayes classifier.

- ○ LDA
- ● QDA
- ○ Naïve Bayes
- ○ Ridge Regression

Solution: It is QDA because each class has its own unique covariance matrix.

Note that Gaussian Naive Bayes is a special case of QDA, but where the covariance matrices are diagonal. This is therefore not Gaussian Naive Bayes.

Problem 11.



The picture above shows the contour lines of a 2-dimensional Gaussian. One of the below options is the Gaussian's covariance matrix. Which is it?

- ○ $\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$
- ● $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$

○ $\begin{pmatrix} 0 & 3 \\ 2 & 0 \end{pmatrix}$

○ $\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$

○ $\begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}$

Solution: The covariance matrix should be diagonal, since the contours are axis-aligned ellipses. We can distinguish between the first two options because the variance in the x direction is smaller than the variance in the y direction.

Problem 12.

The table below collects data on whether or not San Diego Search and Rescue needed to perform a rescue on 9 days last year. For each day, it is listed whether it was during the summer, if it was on the weekend, and whether it was raining.

Day	Is Summer?	Is Weekend?	Is Raining?	Rescue
0	False	False	True	True
1	False	True	True	False
2	False	True	False	False
3	True	True	False	True
4	True	True	False	True
5	True	False	True	False
6	False	False	True	False
7	True	False	False	True
8	True	True	False	True

Use Naive Bayes to predict whether there is a rescue on a day that 1) is not summer; 2) is not a weekend; and 3) is not raining.

Solution: No rescue.

We compute and compare:

$$\underbrace{\mathbb{P}(\text{Not Summer} \,|\, \text{Rescue})}_{1/5} \cdot \underbrace{\mathbb{P}(\text{Not Weekend} \,|\, \text{Rescue})}_{2/5} \cdot \underbrace{\mathbb{P}(\text{Not Raining} \,|\, \text{Rescue})}_{4/5} \cdot \underbrace{\mathbb{P}(\text{Rescue})}_{5/9} = \frac{8}{225}$$

and

$$\underbrace{\mathbb{P}(\text{Not Summer} \,|\, \text{No Rescue})}_{3/4} \cdot \underbrace{\mathbb{P}(\text{Not Weekend} \,|\, \text{No Rescue})}_{2/4} \cdot \underbrace{\mathbb{P}(\text{Not Raining} \,|\, \text{No Rescue})}_{1/4} \cdot \underbrace{\mathbb{P}(\text{No Rescue})}_{4/9} = \frac{6}{144}$$

Since 6/144 is larger than 8/225, we predict No Rescue.

We have:

Problem 13.

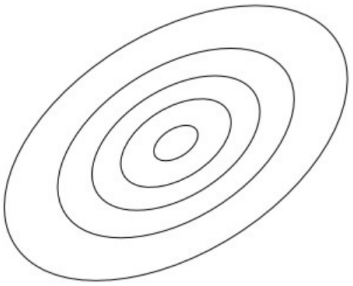
Suppose that 32% of cars in a parking lot are red. Furthermore, suppose that 12% of cars in the parking lot are both red and have four doors. What is the probability that, given a randomly-selected car is red, that is has four doors?

Solution: We use the definition of conditional probability:

$$\begin{aligned}\mathbb{P}(\text{Four Doors} \,|\, \text{Red}) &= \frac{\mathbb{P}(\text{Red and Four Doors})}{\mathbb{P}(\text{Red})} \\ &= \frac{12\%}{32\%} \\ &= 37.5\%\end{aligned}$$

Problem 14.

The figure below shows the contour lines of a 2-dimensional Gaussian. Which one of the following is true? Mark all that apply.



- ■ It has non-zero entries off of the off-diagonal.
- ■ All entries are positive.
- □ All off-diagonal entries are zero, and all diagonal entries are the same.
- □ All off-diagonal entries are zero, and all diagonal entries are different.

Problem 15.

Let X be the number of books sold at the UCSD bookstore in a given week, and let Y be the number of books sold at the UCLA bookstore in the same week. Are X and Y independent or dependent as random variables?

Solution: Dependent. It's likely that busy weeks for UCLA (start of the quarter, holidays, etc.) are also busy at UCSD. Therefore, knowing the number of books sold at UCLA changes our belief in the number of books sold at UCSD.

Problem 16.

Suppose you have two, 6-sided dice, each labeled with the numbers 1 through 6. You roll both dice; let X be the number on the first die, and let Y be the number on the second. Are X and Y independent or dependent?

Solution: Independent. Knowing the result of the first die gives us no information about the result of the second.

Problem 17.

Suppose you have two, 6-sided dice, each labeled with the numbers 1 through 6. You roll both dice; let X be the number on the first die, and let Y be the number on the second.

Are X and Y conditionally independent given that their sum is odd?

Solution: No. If I know that the sum is odd, knowing X in addition gives me information about Y. For example, if I am told that X is 2, then Y must be an odd number.

Because knowing X changes my beliefs about the distribution of Y, $\mathbb{P}(Y \mid X, \text{ sum is odd }) \neq \mathbb{P}(Y \mid \text{ sum is odd })$, and so they are not conditionally independent.