Define the polynomial feature map $\vec{\phi}(\vec{x}) = (1, x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2)^T$. Let $\vec{x} = (1, -1)^T$ and $\vec{x}' = (2, 0)^T$. What is $\vec{\phi}(\vec{x}) \cdot \vec{\phi}(\vec{x}')$?

Hint: there is an easy way, and a hard way, to do this problem.

From Lecture 10, we know that this feature map has a kernel, specifically the polynomial kernel, where $\kappa(\vec{x}, \vec{x}') = (1 + \vec{x} \cdot \vec{x}')^2$. So, $\kappa(\vec{x}, \vec{x}') = \kappa((1, -1)^T, (2, 0)^T) = (1 + 2)^3 = 9$.

Let

- $\vec{x}^{(1)} = (1, 2, 0)^T$
- $\vec{x}^{(2)} = (-1, -1, -1)^T$
- $\vec{x}^{(3)} = (2, 2, 0)^T$
- $\vec{x}^{(4)} = (0, 2, 0)^T$

Suppose a prediction function $H(\vec{x})$ is trained using kernel ridge regression on the data above using the kernel $\kappa(\vec{x}, \vec{x}') = (1 + \vec{x} \cdot \vec{x}')^2$ and regularization parameter $\lambda = 3$. Suppose that $\vec{\alpha} = (1, 0, -1, 2)^T$ is the solution of the dual problem.

Let $\vec{x} = (0, 1, 0)^T$ be a new point. What is $H(\vec{x})$?

To make a prediction using kernels, we can evaluate a weighted sum of kernel evaluations,

$$
\begin{aligned}
H(\vec{x}) &= \sum_{i=1}^{n} \alpha_i^* \kappa(\vec{x}^{(i)}, \vec{x}) \\
&= (1 + \begin{bmatrix}1\\2\\0\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\end{bmatrix})^2 + 0 - (1 + \begin{bmatrix}2\\2\\0\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\end{bmatrix})^2 + 2(1 + \begin{bmatrix}0\\2\\0\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\end{bmatrix})^2 \\
&= (1 + 2)^2 + 0 - (1 + 2)^2 + 2(1 + 2)^2 \\
&= 18
\end{aligned}
$$

An SVM $H(\vec{x})$ with Gaussian kernel is trained with parameter $\gamma = 0.25$. The support vectors (in original data space) are:

- $\vec{x}^{(1)} = (1, 3, 1)^T$
- $\vec{x}^{(2)} = (0, 0, -1)^T$
- $\vec{x}^{(3)} = (-1, -1, 1)^T$

The solution to the dual problem is $\vec{\alpha} = (\frac{1}{2}, 0, -\frac{1}{4})^T$.

Let $\vec{z} = (1, 1, 1)^T$. What is $H(\vec{z})$? Report your answer to two decimal places.

Recall that the Gaussian kernel is $\kappa(\vec{x}, \vec{x}') = e^{-\gamma\|\vec{x}-\vec{x}'\|^2}$, so our prediction evaluates to

$$\begin{aligned} H(\vec{z}) &= \frac{1}{2}(e^{-0.25\|(0,2,0)^T\|^2}) + 0 + (-\frac{1}{4})(e^{-0.25\|(2,2,0)^T\|^2}) \\ &= \frac{1}{2e} - \frac{1}{4e^2} \\ &= 0.15 \end{aligned}$$

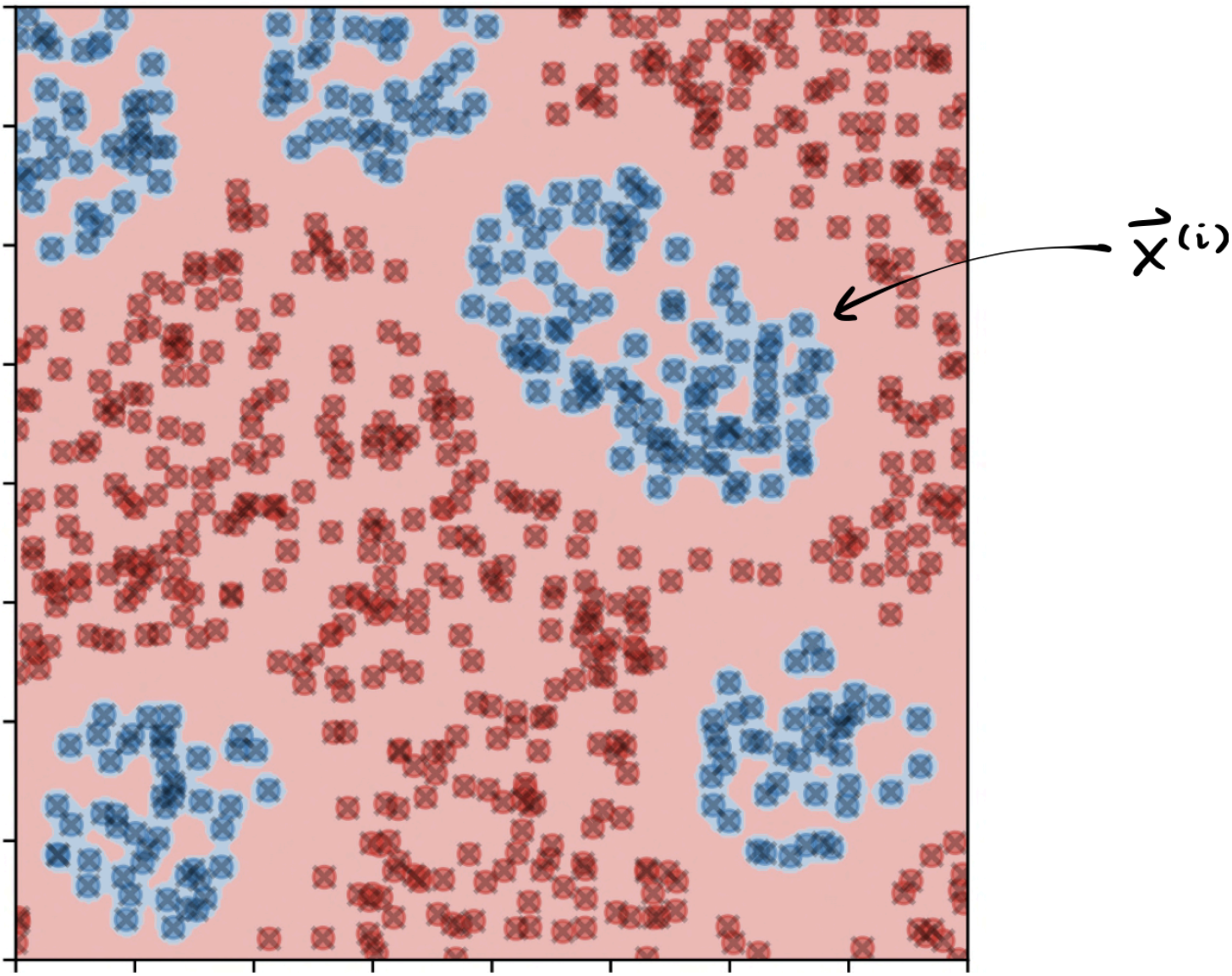
A Gaussian kernel SVM with width parameter γ is trained on a data set and found to overfit. Remember that the RBF kernel is: $\kappa(\vec{x}, \vec{x}') = e^{-\gamma\|\vec{x}-\vec{x}'\|^2}$.

Which of the below should reduce overfitting?

(x) Decreasing γ and re-train
( ) Increasing γ and re-train

Since γ controls the scale of the Gaussians, we would want to increase the width of the Gaussians to reduce overfitting; this translates to a decrease in γ.

A Gaussian kernel SVM is trained on a data set using a very large γ. The result is shown below:



Consider the training point $\vec{x}^{(i)}$ labeled above (the arrow is pointing to the nearest blue support vector). If the blue points have label 1, and the red points have label -1, what will be the learned weight α_i corresponding to this training data point, approximately?

You may assume that γ is as large as you'd like.

First, we know that the value of $H(\vec{x}^{(i)})$ is exactly 1. This is because the data point being pointed to is 1) blue (and blue points have label +1) and 2) it's a support vector, meaning that (by definition) $H(\vec{x}^{(i)}) = y_i = 1$.

Next, remember that for kernel methods $H(\vec{x}) = \sum_{j=1}^{n} \alpha_j \kappa(\vec{x}, \vec{x}^{(j)})$. In this case, κ is the Gaussian kernel, and it measures the "similarity" of its arguments. If $\vec{x}$ and $\vec{x}j$ are "far" away (as controlled by γ), then κ is close to zero; if they are close, κ is apparoximately 1.

In this case, γ is very large, and so $\kappa(\vec{x}, \vec{x}^{(j)})$ is basically zero for any $\vec{x}$ that isn't very close to $\vec{x}^{(j)}$. This means that the only term of $H(\vec{x}^{(i)}) = \sum_{j=1}^{n} \alpha_j \kappa(\vec{x}^{(i)}, \vec{x}^{(j)})$ that *isn't* zero is the term where $j = i$, and that

term is simply α_i. In other words, $H(\vec{x}^{(i)}) \approx \alpha_i$. Since we know that $H(\vec{x}^{(i)} = 1$, this means that $\alpha_i \approx 1$.

The informal explanation is that the surface of H has a bunch of canyons and hills, each one centered at a support vector. The support vector's corresponding α_i controls the height of the hill/canyon. When γ is very large, these hills become very narrow (they are spikes). At $\vec{x}^{(i)}$, the height of the surface is determined almost entirely by the spike located at $\vec{x}^{(i)}$, which is α_i; to make the height 1, α_i should also be 1.

For the parts of this problem, consider the joint distribution shown below:

	$Y = 0$	$Y = 1$
$X = 0$	0%	0%
$X = 1$	5%	0%
$X = 2$	10%	5%
$X = 3$	15%	15%
$X = 4$	5%	20%
$X = 5$	0%	15%
$X = 6$	0%	10%

What is $\mathbb{P}(X = 2)$? State your answer as a decimal number between 0 and 1.

$$\begin{aligned}\mathbb{P}(X = 2) &= \mathbb{P}(X = 2, Y = 0) + \mathbb{P}(X = 2, Y = 1) \\ &= 0.1 + 0.05 \\ &= 0.15\end{aligned}$$

What is $\mathbb{P}(Y = 1)$? State your answer as a decimal number between 0 and 1.

$$\begin{aligned}\mathbb{P}(Y=1) &= \mathbb{P}(X=0,Y=1) + \mathbb{P}(X=1,Y=1) \\ &\quad + \mathbb{P}(X=2,Y=1) + \mathbb{P}(X=3,Y=1) \\ &\quad + \mathbb{P}(X=4,Y=1) + \mathbb{P}(X=5,Y=1) \\ &\quad + \mathbb{P}(X=6,Y=1) \\ &= 0 + 0 + 0.05 + 0.15 + 0.2 + 0.15 + 0.1 \\ &= 0.65\end{aligned}$$

What is $\mathbb{P}(X=2 \,|\, Y=1)$? State your answer as a decimal number between 0 and 1.

$$\mathbb{P}(X=2|Y=1) = \frac{\mathbb{P}(X=2,Y=1)}{\mathbb{P}(Y=1)} = \frac{0.05}{0.65} \approx 0.077$$

What is $\mathbb{P}(Y=1 \,|\, X \leq 3)$? State your answer as a decimal number between 0 and 1.

$$\begin{aligned}\mathbb{P}(Y=1|X\leq 3) &= \frac{\mathbb{P}(X=0,Y=1) + \mathbb{P}(X=1,Y=1) + \mathbb{P}(X=2,Y=1) + \mathbb{P}(X=3,Y=1)}{\mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) + \mathbb{P}(X=3)} \\ &= \frac{0+0+0.05+0.15}{0+0.05+0.15+0.3} \\ &= \frac{0.2}{0.5} \\ &= 0.4\end{aligned}$$

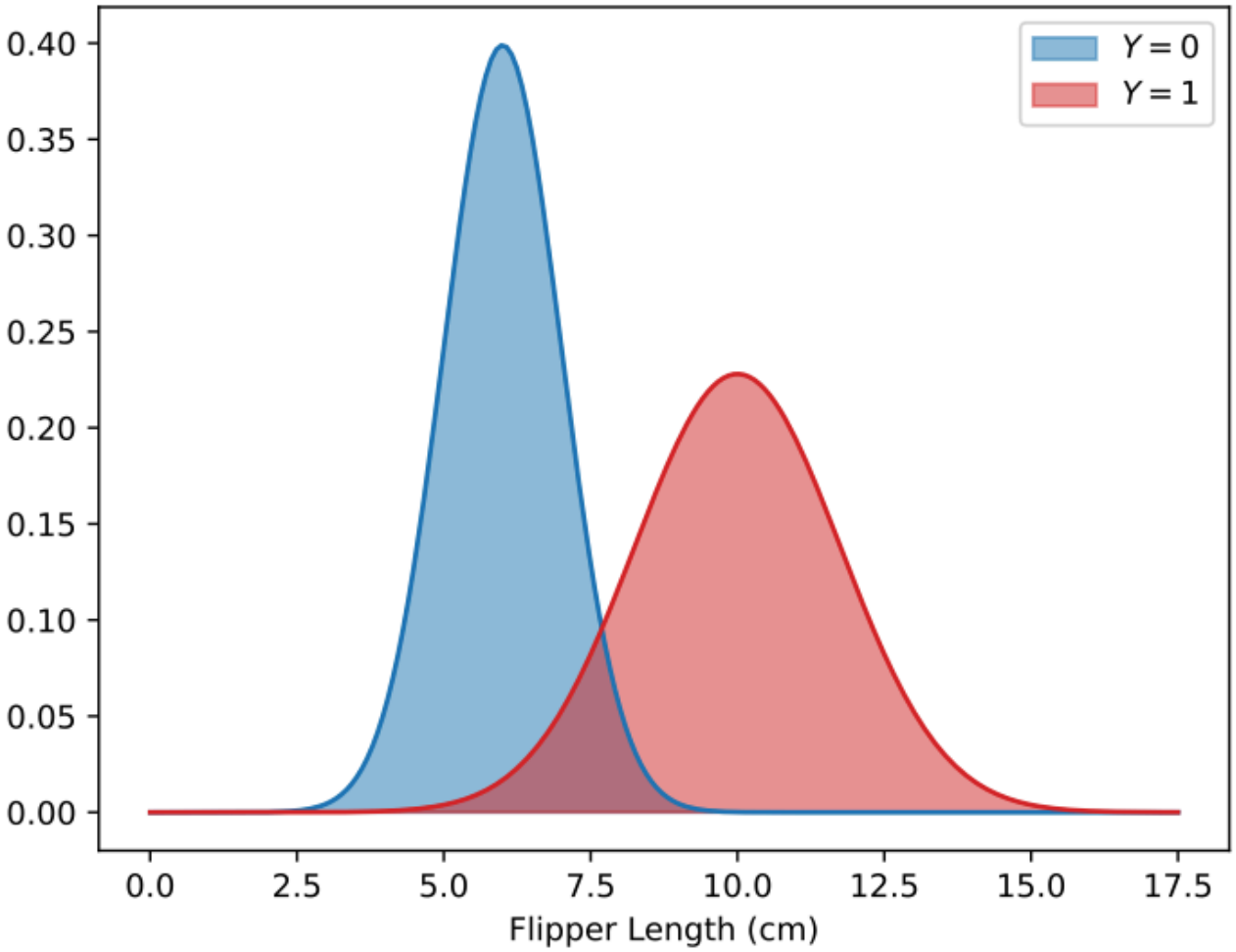
Suppose f is a probability density function (pdf).

True or False: it must be the case that $0 \leq f(x) \leq 1$ for all x.

( ) True
(X) False

The probability density function is a **density** function. This means that the area under the curve, i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$, but the height of the function, i.e. the function value $f(x)$ will depend on the width of the interval, which can be greater than one.

Shown below are the class-conditional densities for penguin flipper length; that is, the densities $p(x \mid Y = 0)$ and $p(x \mid Y = 1)$.



A new penguin is observed with a flipper length of 10cm. What does the Bayes classifier predict for the class of this new penguin?

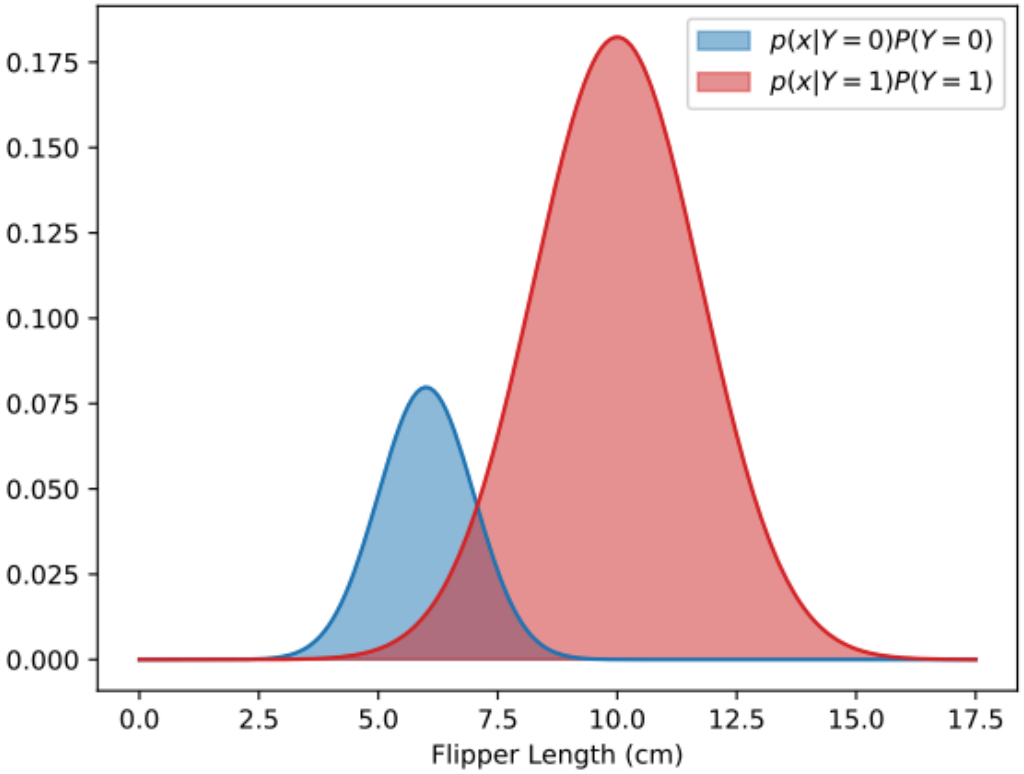
( ) Class Y = 0
( ) Class Y = 1
(X) There is not enough information; we also must know $\mathbb{P}(Y = 1)$ and therefore $\mathbb{P}(Y = 0)$ in order to make a prediction.

Recall that the Bayes classification rule predicts class 1 if

$$p(X = x|Y = 1)\mathbb{P}(Y = 1) > p(X = x|Y = 0)\mathbb{P}(Y = 0).$$

Since we don't have the class "prior" probabilities (we only have the class-conditional densities), there is not enough information to make a prediction.

Shown below are the class-conditional densities for penguin flipper length weighted by the class "prior" probabilities; that is, the functions $p(x \mid Y = 0)\mathbb{P}(Y = 0)$ and $p(x \mid Y = 1)\mathbb{P}(Y = 1)$.

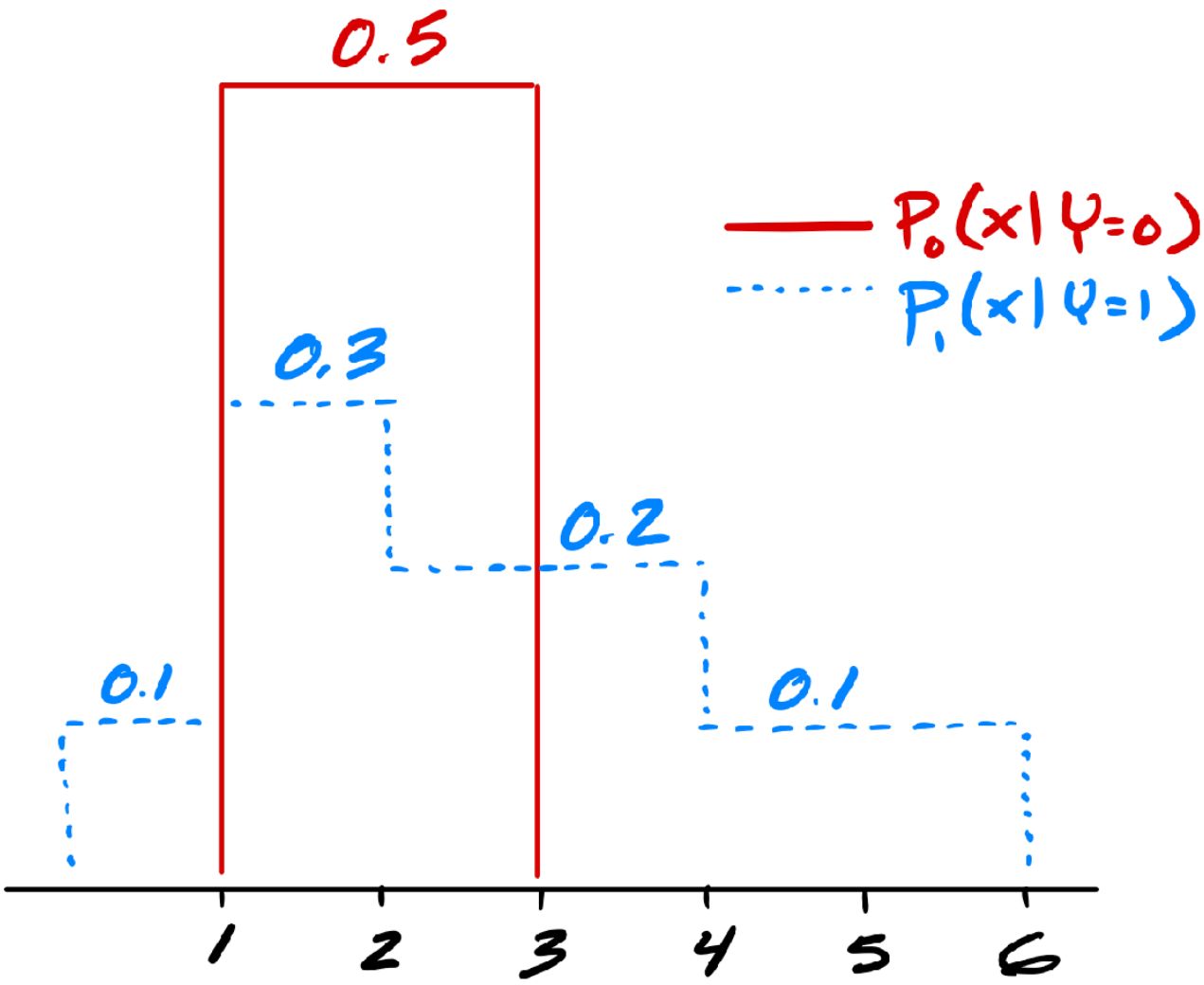


A new penguin is observed with a flipper length of 10cm. What does the Bayes classifier predict for the class of this new penguin?

( ) Class Y = 0
(X) Class Y = 1
( ) There is not enough information; we also must know the marginal density $p(x)$ in order to make a prediction.

Since the red curve (weighted class-conditional density of $Y = 1$) has a higher value than that of the blue curve (weighted class-conditional density of $Y = 0$) at a flipper length of 10 cm, the Bayes classifier would predict class $Y = 1$.

Shown 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 variable X for two classes: $Y = 0$ (the solid line) and $Y = 1$ (the dashed line).
You may assume that both densities are piecewise constant.



Suppose $\mathbb{P}(Y=1)=0.5$ and $\mathbb{P}(Y=0)=0.5$. What is the prediction of the Bayes classifier at $x=1.5$?

(x) Class 0 (the solid class)
( ) Class 1 (the dashed class)

According to Bayes classification rule, we need to check which of the two are larger: $p_0(x|Y=0)\mathbb{P}(Y=0)$ or $p_1(x|Y=1)\mathbb{P}(Y=1)$.

Since $\mathbb{P}(Y=0)=\mathbb{P}(Y=1)$, we can essentially ignore these class "prior" probabilities, and just compare the two conditional densities.

Observe that at $x=1.5$, we have that

$$p_0(1.5|Y=0)=0.5>0.3=p_1(1.5|Y=1).$$

Thus, the Bayes classifier predicts Class 0.

Suppose $\mathbb{P}(Y=1)=0.5$ and $\mathbb{P}(Y=0)=0.5$. What is the Bayes error with respect to this distribution?

Recall that the Bayes error can be calculated as $\mathbb{P}(\text{error}) = \mathbb{P}(Y \text{ is actually } 1)p_1(\text{predict } 0|Y=1) + \mathbb{P}(Y \text{ is actually } 0)p_0(\text{predict } 1|Y=0)$.

In this scenario, we only have one case to consider: when the Bayes classifier predicts 0 when Y is actually 1. This is in the interval $[1,3]$. So, Bayes error is

$$\begin{aligned}\mathbb{P}(\text{error}) &= \mathbb{P}(Y \text{ is actually } 1)p_1(\text{predict } 0|Y=1) \\ &= 0.5(0.3+0.2) \\ &= 0.25\end{aligned}$$

Now suppose $\mathbb{P}(Y=1)=0.7$ and $\mathbb{P}(Y=0)=0.3$. What is the prediction of the Bayes classifier at $x=1.5$?

( ) Class 0 (the solid class)
(X) Class 1 (the dashed class)

We repeat the same steps as Q10.1, but now we have different class "prior" probabilities; this means we must take them into account when comparing the two classes.

Observe that at $x=1.5$, we have

$$p_0(1.5|Y=0)\mathbb{P}(Y=0) = 0.5 \cdot 0.3 = 0.15 < 0.21 = 0.3 \cdot 0.7 = p_1(1.5|Y=1)\mathbb{P}(Y=1).$$

Thus, the Bayes classifier predicts Class 1.

Now suppose $\mathbb{P}(Y=1)=0.7$ and $\mathbb{P}(Y=0)=0.3$. What is the Bayes error with respect to this distribution? Recall that the Bayes error is the probability that the Bayes classifier makes the wrong prediction.

Since we are now comparing the weighted class-conditional densities, the picture will look a bit different.

Again, we calculate

$$\mathbb{P}(\text{error}) = \mathbb{P}(Y \text{ is actually } 1)p_1(\text{predict } 0|Y=1) + \mathbb{P}(Y \text{ is actually } 0)p_0(\text{predict } 1|Y=0).$$

In this scenario, we see that both cases are possible. In particular, the Bayes classifier predicts 0 when Y is actually 1 in the interval $[2, 3]$ and predicts 1 when Y is actually 0 in the interval $[1, 2]$. So, Bayes error is $\mathbb{P}(\text{error}) = 0.7 \cdot 0.2 + 0.3 \cdot 0.5 = 0.29$.

Suppose a particular probability distribution has the property that, whenever data are sampled from the distribution, the sampled data are guaranteed to be linearly separable. True or False: the Bayes error with respect to this distribution is 0%.

(x) True
( ) False

If the data are guaranteed to be linearly separable, then there is a line that will have zero classification error on any data set drawn from the distribution, no matter how large. In other words, the probability of this classifier making an error is 0. The Bayes error by definition the smallest possible error probability of any classifier, and you can't have an error smaller than zero, so this is true.

Now consider a different probability distribution. Suppose the Bayes classifier achieves an error rate of 0% on this distribution. True or False: given a finite data set sampled from this distribution, the data must be linearly separable.

( ) True
(x) False

Consider a probability distribution that randomly generates a point in two dimensions and assigns a color based on the following: if it is in the northeast or southwest quadrant, it is red; otherwise, blue.

A data set generated from this distribution is (generally) not linearly separable, as drawing a picture will show.

On the other hand, there is a classification rule that has zero error. Namely: if the point is in the northeast or southwest quadrants, predict red; otherwise predict blue. This rule never makes a mistake, and so the Bayes error with respect to this distribution is zero, but the data are not linearly separable in general.

Informally-speaking, Bayes error measure the amount of "ambiguity" there can be about a point's label. For this generating process, there's no ambiguity: a point's label is entirely determined by where it is located; there's no "overlap" between red and blue. That doesn't mean that the data drawn from the distribution are linearly separable, though.