# Midterm Exam

CMU 10-601: Machine Learning (Spring 2016)

Feb. 29, 2016

**Name:** _____

**Andrew ID:** _____

## START HERE: Instructions

- This exam has 16 pages and 5 Questions (page one is this cover page). Check to see if any pages are missing. Enter your name and Andrew ID above.
- You are allowed to use one page of notes, front and back.
- Electronic devices are not acceptable.
- Note that the questions vary in difficulty. Make sure to look over the entire exam before you start and answer the easier questions first.

| Question | Point | Score |
|---|---|---|
| 1 | 20 | |
| 2 | 20 | |
| 3 | 20 | |
| 4 | 20 | |
| 5 | 20 | |
| Extra Credit | 14 | |
| Total | 114 | |

# 1 Naive Bayes, Probability, and MLE [20 pts. + 2 Extra Credit]

## 1.1 Naive Bayes

You are given a data set of 10,000 students with their sex, height, and hair color. You are trying to build a classifier to predict the sex of a student, so you randomly split the data into a training set and a testing set. Here are the specifications of the data set:

- sex $\in$ {male,female}
- height $\in$ [0,300] centimeters
- hair $\in$ {brown, black, blond, red, green}
- 3240 men in the data set
- 6760 women in the data set

Under the assumptions necessary for Naive Bayes (not the distributional assumptions you might naturally or intuitively make about the dataset) answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

(a) [2 pts.] **T or F:** As height is a continuous valued variable, Naive Bayes is not appropriate since it cannot handle continuous valued variables.

(b) [2 pts.] **T or F:** Since there is not a similar number of men and women in the dataset, Naive Bayes will have high test error.

(c) [2 pts.] **T or F:** $P(\texttt{height}|\texttt{sex},\texttt{hair}) = P(\texttt{height}|\texttt{sex})$.

(d) [2 pts.] **T or F:** $P(\texttt{height},\texttt{hair}|\texttt{sex}) = P(\texttt{height}|\texttt{sex})P(\texttt{hair}|\texttt{sex})$.

## 1.2 Maximum Likelihood Estimation (MLE)

Assume we have a random sample that is Bernoulli distributed $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. We are going to derive the MLE for $\theta$. Recall that a Bernoulli random variable $X$ takes values in $\{0, 1\}$ and has probability mass function given by

$$P(X; \theta) = \theta^X (1 - \theta)^{1-X}.$$

(a) [2 pts.] Derive the likelihood, $L(\theta; X_1, \ldots, X_n)$.

(b) [2 pts.] Derive the following formula for the log likelihood:

$$\ell(\theta; X_1, \ldots, X_n) = \left( \sum_{i=1}^{n} X_i \right) \log(\theta) + \left( n - \sum_{i=1}^{n} X_i \right) \log(1 - \theta).$$

(c) **Extra Credit:** [2 pts.] Derive the following formula for the MLE: $\hat{\theta} = \frac{1}{n} (\sum_{i=1}^{n} X_i)$.

## 1.3 MAP vs MLE

Answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

(a) [2 pts.] **T or F:** In the limit, as $n$ (the number of samples) increases, the MAP and MLE estimates become the same.

(b) [2 pts.] **T or F:** Naive Bayes can only be used with MAP estimates, and not MLE estimates.

## 1.4 Probability

Assume we have a sample space $\Omega$. Answer each question with **T** or **F**. **No justification is required.**

(a) [1 pts.] **T or F:** If events $A$, $B$, and $C$ are disjoint then they are independent.

(b) [1 pts.] **T or F:** $P(A|B) \propto \frac{P(A)P(B|A)}{P(A|B)}$. (The sign '$\propto$' means 'is proportional to')

(c) [1 pts.] **T or F:** $P(A \cup B) \leq P(A)$.

(d) [1 pts.] **T or F:** $P(A \cap B) \geq P(A)$.

# 2 Errors, Errors Everywhere [20 pts.]

## 2.1 True Errors

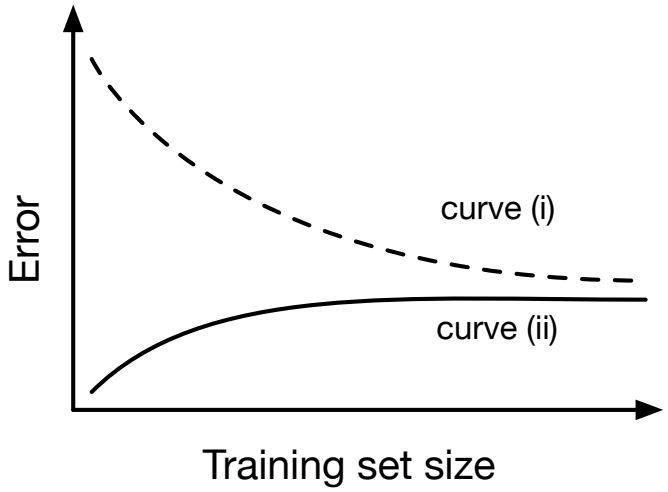
Consider a classification problem on $\mathbb{R}^d$ with distribution $D$ and target function $c^* : \mathbb{R}^d \to \{\pm 1\}$. Let $S$ be an iid sample drawn from the distribution $D$. Answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

(a) [4 pts.] **T or F:** The true error of a hypothesis $h$ can be lower than its training error on the sample $S$.

(b) [4 pts.] **T or F:** Learning theory allows us to determine with 100% certainty the true error of a hypothesis to within any $\epsilon > 0$ error.

## 2.2 Training Sample Size

In this problem, we will consider the effect of training sample size $n$ on a logistic regression classifier with $d$ features. The classifier is trained by optimizing the conditional log-likelihood. The optimization procedure stops if the estimated parameters perfectly classify the training data or they converge. The following plot shows the general trend for how the training and testing error change as we increase the sample size $n = |S|$. Your task in this question is to analyze this plot and identify which curve corresponds to the training and test error. Specifically:




(a) [8 pts.] Which curve represents the training error? **Please provide 1–2 sentences of justification**.

(b) [4 pt.] In one word, what does the gap between the two curves represent?

# 3 Linear and Logistic Regression [20 pts. + 2 Extra Credit]

## 3.1 Linear regression

Given that we have an input $x$ and we want to estimate an output $y$, in linear regression we assume the relationship between them is of the form $y = wx + b + \epsilon$, where $w$ and $b$ are real-valued parameters we estimate and $\epsilon$ represents the noise in the data. When the noise is Gaussian, maximizing the likelihood of a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to estimate the parameters $w$ and $b$ is equivalent to minimizing the squared error:

$$\arg\min_w \sum_{i=1}^{n} (y_i - (wx_i + b))^2.$$

Consider the dataset $S$ plotted in Fig. 1 along with its associated regression line. For each of the altered data sets $S^{\text{new}}$ plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

| Dataset | (a) | (b) | (c) | (d) | (e) |
|---|---|---|---|---|---|
| Regression line | | | | | |

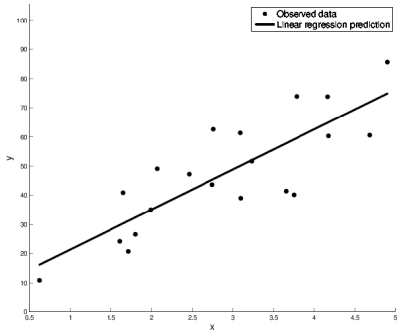
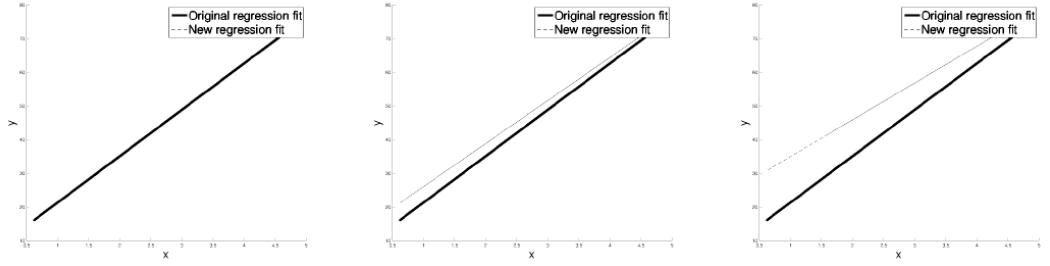
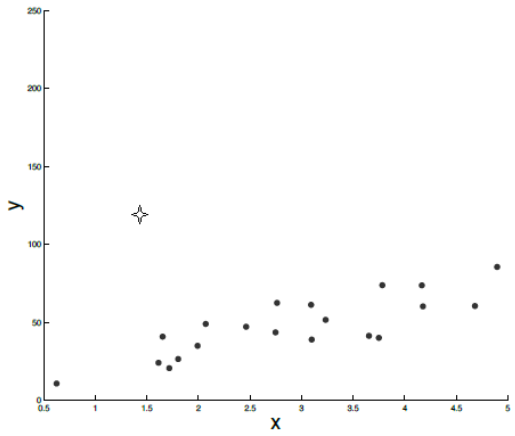



Figure 1: An observed data set and its associated regression line.

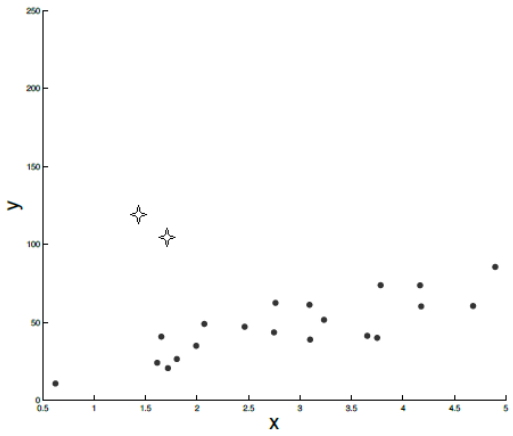



(a) Old and new regression lines. (b) Old and new regression lines. (c) Old and new regression lines.

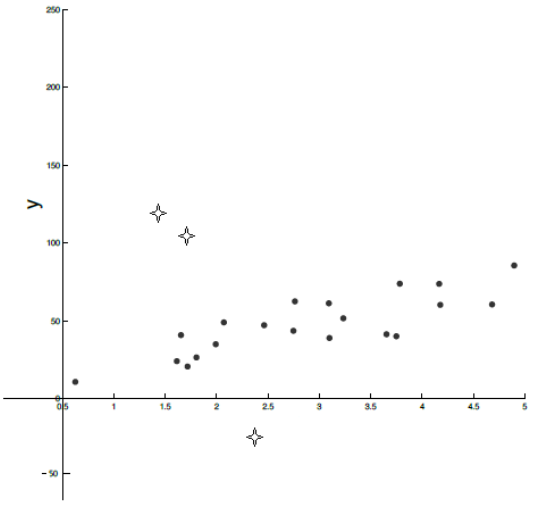
Figure 2: New regression lines for altered data sets $S^{\text{new}}$.



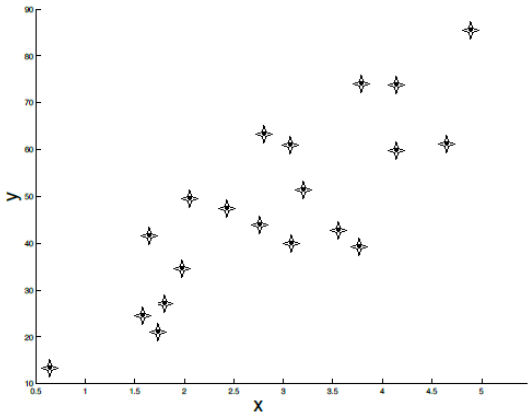

(a) Adding one outlier to the original data set.



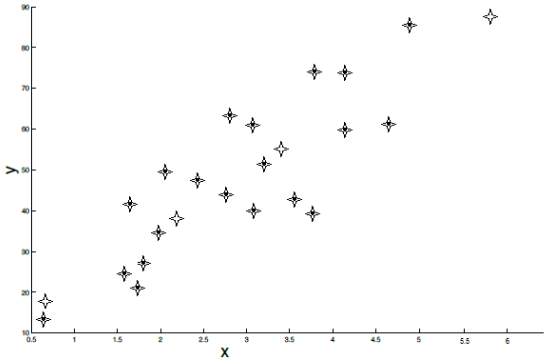

(b) Adding two outliers to the original data set.




(c) Adding three outliers to the original data set. Two on one side and one on the other side.




(d) Duplicating the original data set.




(e) Duplicating the original data set and adding four points that lie on the trajectory of the original regression line.

Figure 3: New data set $S^{\text{new}}$.

## 3.2 Logistic regression

Given a training set $\{(x_i, y_i), i = 1, \ldots, n\}$ where $x_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{0, 1\}$ is a binary label, we want to find the parameters $\hat{w}$ that maximize the likelihood for the training set, assuming a parametric model of the form

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w^T x)}.$$

The conditional log likelihood of the training set is

$$\ell(w) = \sum_{i=1}^{n} y_i \log p(y_i, |x_i; w) + (1 - y_i) \log(1 - p(y_i, |x_i; w)),$$

and the gradient is

$$\nabla \ell(w) = \sum_{i=1}^{n} (y_i - p(y_i|x_i; w))x_i.$$

(a) [5 pts.] Is it possible to get a closed form for the parameters $\hat{w}$ that maximize the conditional log likelihood? How would you compute $\hat{w}$ in practice?

(b) [5 pts.] What is the form of the classifier output by logistic regression?

(c) [2 pts.] **Extra Credit:** Consider the case with binary features, i.e, $x \in \{0, 1\}^d \subset \mathbb{R}^d$, where feature $x_1$ is rare and happens to appear in the training set with only label 1. What is $\hat{w}_1$? Is the gradient ever zero for any finite $w$? Why is it important to include a regularization term to control the norm of $\hat{w}$?

# 4 SVM, Perceptron and Kernels [20 pts. + 4 Extra Credit]

## 4.1 True or False

Answer each of the following questions with **T** or **F** and **provide a one line justification**.

(a) [2 pts.] Consider two datasets $D^{(1)}$ and $D^{(2)}$ where $D^{(1)} = \{(x_1^{(1)}, y_1^{(1)}), ..., (x_n^{(1)}, y_n^{(1)})\}$ and $D^{(2)} = \{(x_1^{(2)}, y_1^{(2)}), ..., (x_m^{(2)}, y_m^{(2)})\}$ such that $x_i^{(1)} \in \mathbb{R}^{d_1}$, $x_i^{(2)} \in \mathbb{R}^{d_2}$. Suppose $d_1 > d_2$ and $n > m$. Then the maximum number of mistakes a perceptron algorithm will make is higher on dataset $D^{(1)}$ than on dataset $D^{(2)}$.

(b) [2 pts.] Suppose $\phi(\mathbf{x})$ is an arbitrary feature mapping from input $\mathbf{x} \in \mathcal{X}$ to $\phi(\mathbf{x}) \in \mathbb{R}^N$ and let $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$. Then $K(\mathbf{x}, \mathbf{z})$ will always be a valid kernel function.

(c) [2 pts.] Given the same training data, in which the points are linearly separable, the margin of the decision boundary produced by SVM will always be greater than or equal to the margin of the decision boundary produced by Perceptron.

## 4.2 Multiple Choice

(a) [3 pt.] If the data is linearly separable, SVM minimizes $\|w\|^2$ subject to the constraints $\forall i, y_i w \cdot x_i \geq 1$. In the linearly separable case, which of the following may happen to the decision boundary if one of the training samples is removed? **Circle all that apply.**

- Shifts toward the point removed
- Shifts away from the point removed
- Does not change

(b) [3 pt.] Recall that when the data are not linearly separable, SVM minimizes $\|w\|^2 + C \sum_i \xi_i$ subject to the constraint that $\forall i$, $y_i w \cdot x_i \geq 1 - \xi_i$ and $\xi_i \geq 0$. Which of the following may happen to the size of the margin if the tradeoff parameter $C$ is increased? **Circle all that apply.**

- Increases
- Decreases
- Remains the same

## 4.3 Analysis

(a) [4 pts.] In one or two sentences, describe the benefit of using the Kernel trick.

(b) [4 pt.] The concept of margin is essential in both SVM and Perceptron. Describe why a large margin separator is desirable for classification.

(c) [4 pts.] **Extra Credit:** Consider the dataset in Fig. 4. Under the SVM formulation in section 4.2(a),

(1) Draw the decision boundary on the graph.

(2) What is the size of the margin?

(3) Circle all the support vectors on the graph.

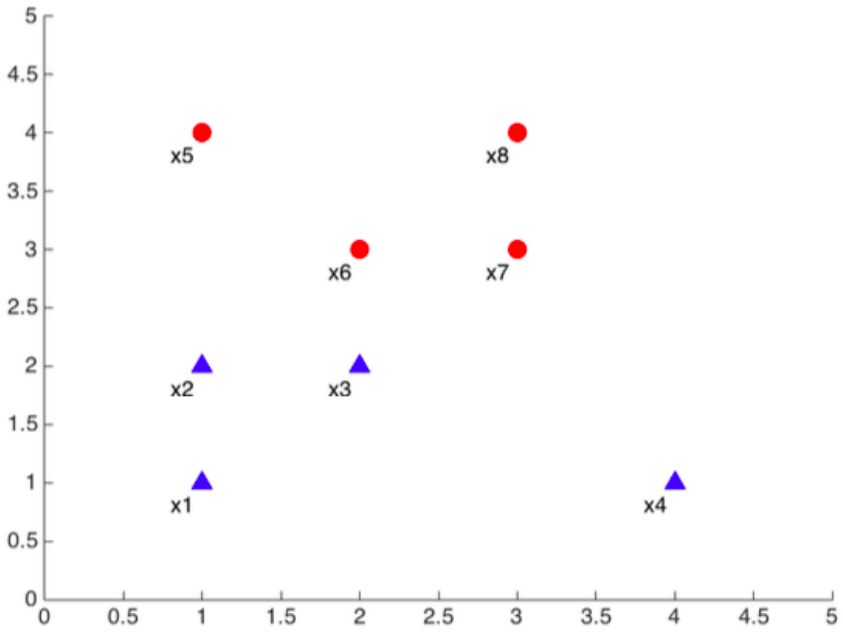



Figure 4: SVM toy dataset

# 5 Learning Theory [20 pts.]

## 5.1 True or False

Answer each of the following questions with **T** or **F** and **provide a one line justification.**

(a) [3 pts.] **T or F**: It is possible to label 4 points in $\mathbb{R}^2$ in all possible $2^4$ ways via linear separators in $\mathbb{R}^2$.

(b) [3 pts.] **T or F**: To show that the VC-dimension of a concept class $H$ (containing functions from $X$ to $\{0, 1\}$) is $d$, it is sufficient to show that there exists a subset of $X$ with size $d$ that can be labeled by $H$ in all possible $2^d$ ways.

(c) [3 pts.] **T or F**: The VC dimension of a finite concept class $H$ is upper bounded by $\lceil \log_2 |H| \rceil$.

(d) [3 pts.] **T or F**: The VC dimension of a concept class with infinite size is also infinite.

(e) [3 pts.] **T or F**: For every pair of classes, $H_1$, $H_2$, if $H_1 \subseteq H_2$ and $H_1 \neq H_2$, then $\text{VCdim}(H_1) < \text{VCdim}(H_2)$ (note that this is a strict inequality).

(f) [3 pts.] **T or F**: Given a realizable concept class and a set of training instances, a consistent learner will output a concept that achieves 0 error on the training instances.

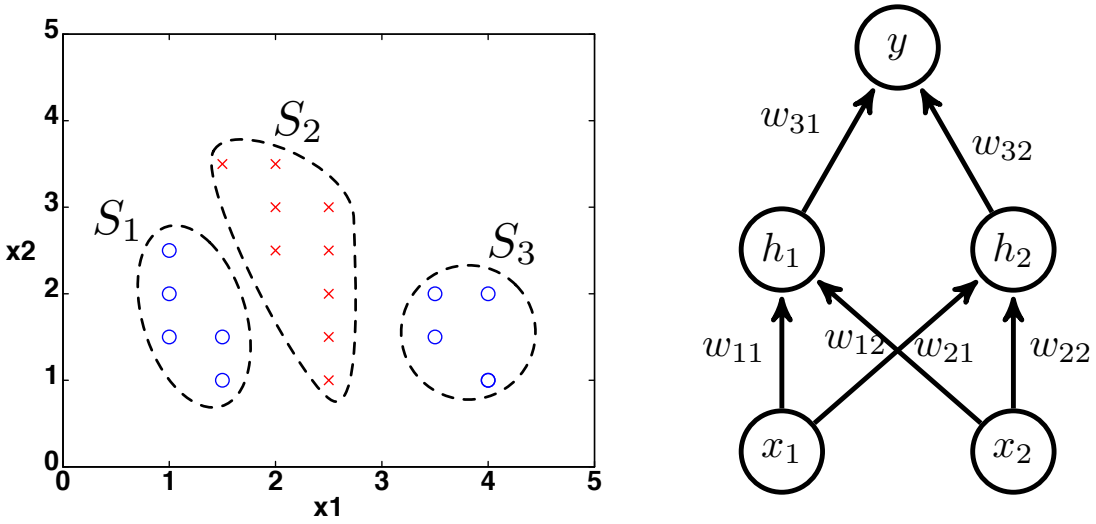
## 5.2 VC dimension

Briefly explain **in 2–3 sentences** the importance of sample complexity and VC dimension in learning with generalization guarantees.

## 6 Extra Credit: Neural Networks [6 pts.]

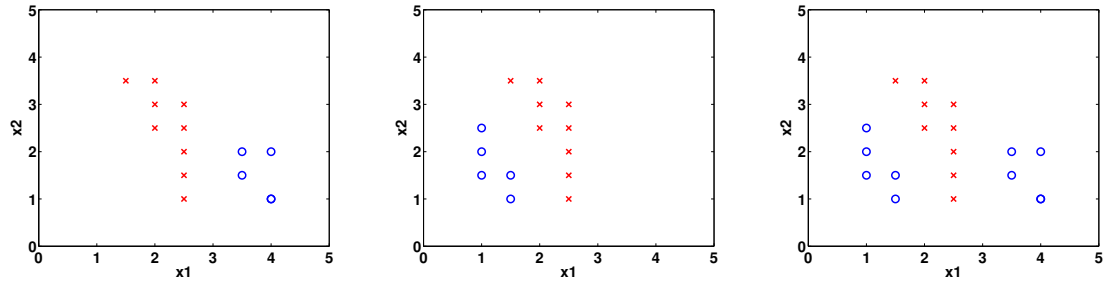
In this problem we will use a neural network to classify the crosses ($\times$) from the circles ($\circ$) in the simple dataset shown in Figure 5a. Even though the crosses and circles are not linearly separable, we can break the examples into three groups, $S_1$, $S_2$, and $S_3$ (shown in Figure 5a) so that $S_1$ is linearly separable from $S_2$ and $S_2$ is linearly separable from $S_3$. We will exploit this fact to design weights for the neural network shown in Figure 5b in order to correctly classify this training set. For all nodes, we will use the threshold activation function

$$\phi(z) = \begin{cases} 1 & z > 0 \\ 0 & z \leq 0. \end{cases}$$




(a) The dataset with groups $S_1$, $S_2$, and $S_3$. (b) The neural network architecture

Figure 5




(a) Set S2 and S3 (b) Set S1 and S2 (c) Set S1, S2 and S3

Figure 6: NN classification.

(a) First we will set the parameters $w_{11}, w_{12}$ and $b_1$ of the neuron labeled $h_1$ so that its output $h_1(x) = \phi(w_{11}x_1 + w_{12}x_2 + b_1)$ forms a linear separator between the sets $S_2$ and $S_3$.

(1) [1 pt.] On Fig. 6a, draw a linear decision boundary that separates $S_2$ and $S_3$.

(2) [1 pt.] Write down the corresponding weights $w_{11}, w_{12}$, and $b_1$ so that $h_1(x) = 0$ for all points in $S_3$ and $h_1(x) = 1$ for all points in $S_2$.

(b) Next we set the parameters $w_{21}, w_{22}$ and $b_2$ of the neuron labeled $h_2$ so that its output $h_2(x) = \phi(w_{21}x_1 + w_{22}x_2 + b_2)$ forms a linear separator between the sets $S_1$ and $S_2$.

(1) [1 pt.] On Fig. 6b, draw a linear decision boundary that separates $S_1$ and $S_2$.

(2) [1 pt.] Write down the corresponding weights $w_{21}, w_{22}$, and $b_2$ so that $h_2(x) = 0$ for all points in $S_1$ and $h_2(x) = 1$ for all points in $S_2$.

(c) Now we have two classifiers $h_1$ (to classify $S_2$ from $S_3$) and $h_2$ (to classify $S_1$ from $S_2$). We will set the weights of the final neuron of the neural network based on the results from $h_1$ and $h_2$ to classify the crosses from the circles. Let $h_3(x) = \phi\big(w_{31}h_1(x) + w_{32}h_2(x) + b_3\big)$.

(1) [1 pt.] Compute $w_{31}, w_{32}, b_3$ such that $h_3(x)$ correctly classifies the entire dataset.

(2) [1 pt.] Draw your decision boundary in Fig. 6c.

Use this page for scratch work