A) Problems from the Textbook

0.1 Learning Exercises

Exercise 1.1

Express each of the following tasks in the framework of learning from data by specifying the input space X, output space Y, target function f: X -> Y, and the specifics of the data set that we will learn from.

(a) Medical diagnosis: A patient walks in with a medical history and some symptoms, and you want to identify the problem.

Arguments: This is a multiclass classification problem, that can be solved by supervised learning, where each class represents a disease. We can use concrete features of patients such as their medical history and symptoms in order to identify a disease.

Answer:

- the input space X is a medical history and symptoms;
- the output space Y is the medical diagnoses (diseases);
- the target function f: X -> Y is a formula to correctly determine the medical diagnosis (disease), based on the medical history and symptoms;
- the data set is structured with concrete features.

(b) Handwritten digit recognition (for example postal zip code recognition for mail sorting).

Arguments: This is a multiclass classification problem with ten classes that represent every digit from 0 to 9. We can use labeled images of handwritten digits in order to train the model. After that we can recognize postal zip code digit by digit.

Answer:

- the input space X is images of handwritten digits from 0 to 9;
- the output space Y is digits from 0 to 9;
- the target function f: X -> Y is a formula to correctly match image of handwritten digit to corresponding digit;
- the data is unstructured with raw features.

(c) Determining if an email is spam or not.

Arguments: This is a binary classification problem, that can be solved by natural language processing. We can detect some keywords in an email in order to determine if it is a spam or not.

Answer:

- the input space X is a set of words;
- the output space Y is True/False;
- the target function f: X -> Y is a formula to correctly distinguish between spam and not spam;
- the data is unstructured, NLP is useful for this problem.

(d) *Predicting how an electric load varies with price, temperature, and day of the week.*

Arguments: This is a regression problem, that can be solved by supervised learning. We can use concrete features such as price, temperature, and weekday in order to predict the amount of electric load.

Answer:

- the input space X price, temperature, and day of the week;
- the output space Y is amount of electric load = R;
- the target function f: X -> Y is a formula to correctly identify the amount of electric load, based on price, temperature, and day of the week;
- the data is structured with concrete features.

(d) A problem of interest: Driver drowsiness detection

Arguments: This is a binary classification problem, that can be solved by supervised learning and CNN. One of the way to solve this problem could be to estimate eye closure. We can detect the eyes of the driver and identify if they are close or open, and if eyes are close longer than some threshold, detect it as drowsiness.

Answer:

- the input space X is images of open and close eyes;
- the output space Y is True/False;
- the target function f: X -> Y is a formula to correctly detect if a driver is drowsy or not;
- the data is raw and unstructured.

Exercise 1.5

Which of the following problems are more suited for the learning approach and which are more suited for the design approach?

(a) Determining the age at which a particular medical test should be performed
Arguments: If we think about this problem in terms of determining the age at which a person is more likely to contract a particular disease and should be recommended to perform a medical test, the learning approach is more suitable. Because have enough data in the medical database.

But if we think in terms of at what age it is allowed to perform a particular test without negative consequences, the design approach is more appropriate. Because we do not have enough data for the new medical test to learn from. But we can determine the age at which this test should be performed based on such medical specifications as contraindications, age restrictions for disease etc.
Answer: Learning Approach/Design Approach

(b) Classifying numbers into primes and non-primes
Arguments: This problem is more suited for the design approach. Because we do know the definition of a prime number. A prime number is a natural number greater than 1 and having no positive divisor other than 1 and itself. Based on this knowledge we can easily solve this problem using algorithmic approach.
Answer: Design Approach

(c) *Detecting potential fraud in credit card charges*
Arguments: This problem is more suited for the learning approach. Because we have enough data of transaction that includes both fraudulent and non-fraudulent ones, there is a pattern, and it is very complicated to correctly detect using a design approach.
Answer: Learning Approach

(d) *Determining the time it would take a falling object to hit the ground*
Arguments: This problem is more suited for the design approach. Because we do know the exact formula for that problem. And we can compute the time by specifying the initial velocity, distance to the ground and acceleration of the falling object.
Answer: Design Approach

(e) *Determining the optimal cycle for traffic lights in a busy intersection*

Arguments: This problem is more suited for the design approach because there is a simple direct relationship between the density of traffic in each direction and the cycle for traffic lights. This relationship can be easily observed by a human who determines the optimal cycle, and the problem can be solved. Although the learning approach also can be applied, I think it is not worthwhile, in this particular problem.

Answer: Design Approach

0.2 Perceptron Learning Algorithm

Exercise 1.3

a) Show that $y(t)w^T(t)x(t) < 0$. [Hint: x(t) is misclassified by w(t)]

Solution:

Since x(t) is misclassified by w(t) the signs of y(t) and $w^T(t)x(t)$ are opposite. The multiplication of two opposite signs gives a negative sign.

- If y(t) = +1, then sign($w^T(t)x(t)$) = -1, meaning that $w^T(t)x(t)$ is negative number; +1*(negative number) < 0, therefore, $y(t)w^T(t)x(t) < 0$;
- If y(t) = -1, then sign($w^T(t)x(t)$) = +1, meaning that $w^T(t)x(t)$ is positive number; -1*(positive number) < 0, therefore, $y(t)w^T(t)x(t) < 0$;

b) Show that $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$

Solution:

Substitute w(t+1) by w(t) + y(t)x(t)

$y(t)[w(t) + y(t)x(t)]^Tx(t) > y(t)w^T(t)x(t)$

$y(t)w^T(t)x(t) + y(t)y(t)x^T(t)x(t) > y(t)w^T(t)x(t)$

$y^2(t)x^T(t)x(t)$ is a positive number, therefore $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$

c) As far as classifying x(t) is concerned, argue that the move from w(t) to w(t + 1) is a move ' in the right direction'

Solution:

The dot product $w^T(t)x(t)$ can be expressed as $||w^T||||x||\cos(\alpha)$, where α is an angle between w and x. Since the lengths are always positive, the sign of $w^T(t)x(t)$ depends on a cosine of the angle between w and x. If the sign of y(t) an $w^T(t)x(t)$ does not match, the update rule, w(t+1) = w(t) + y(t)x(t) will create a new vector that will change the angle in a right direction.

0.3 Experiments with Perceptron Learning Algorithm

See the source code in DenisKimHomework1.ipynb

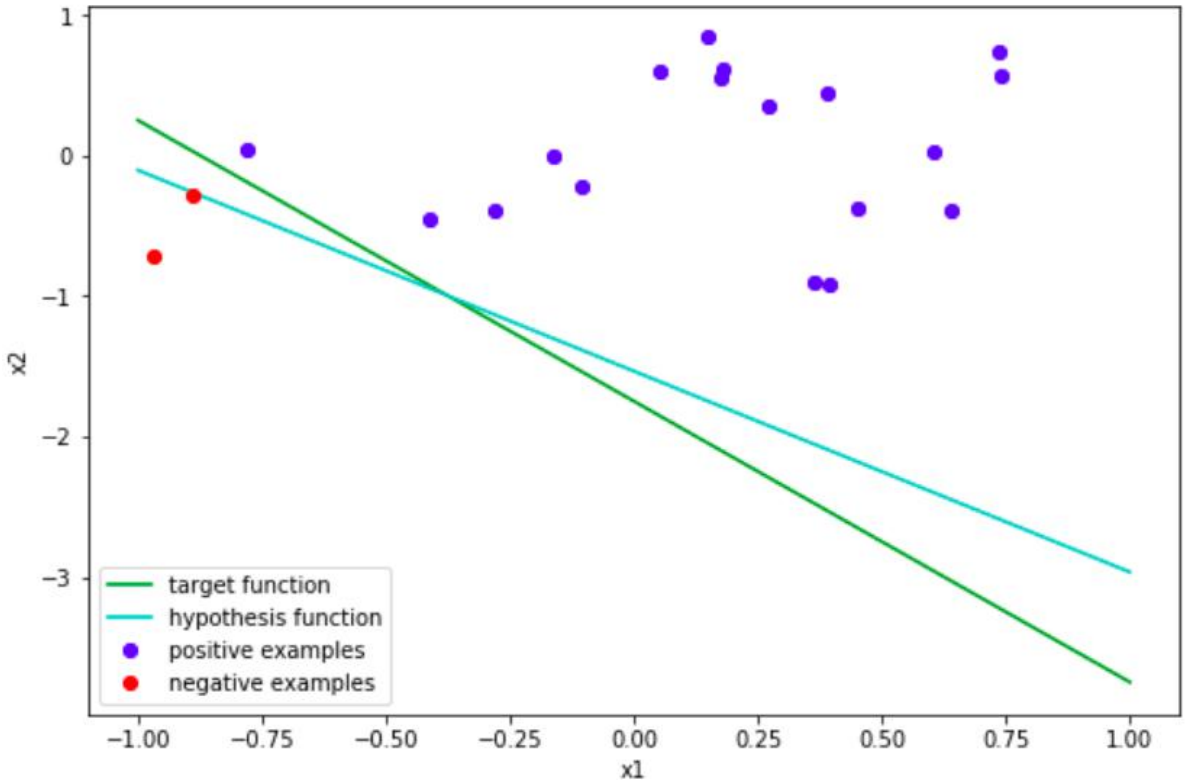

Figure1: The graph for the first data set of size 20

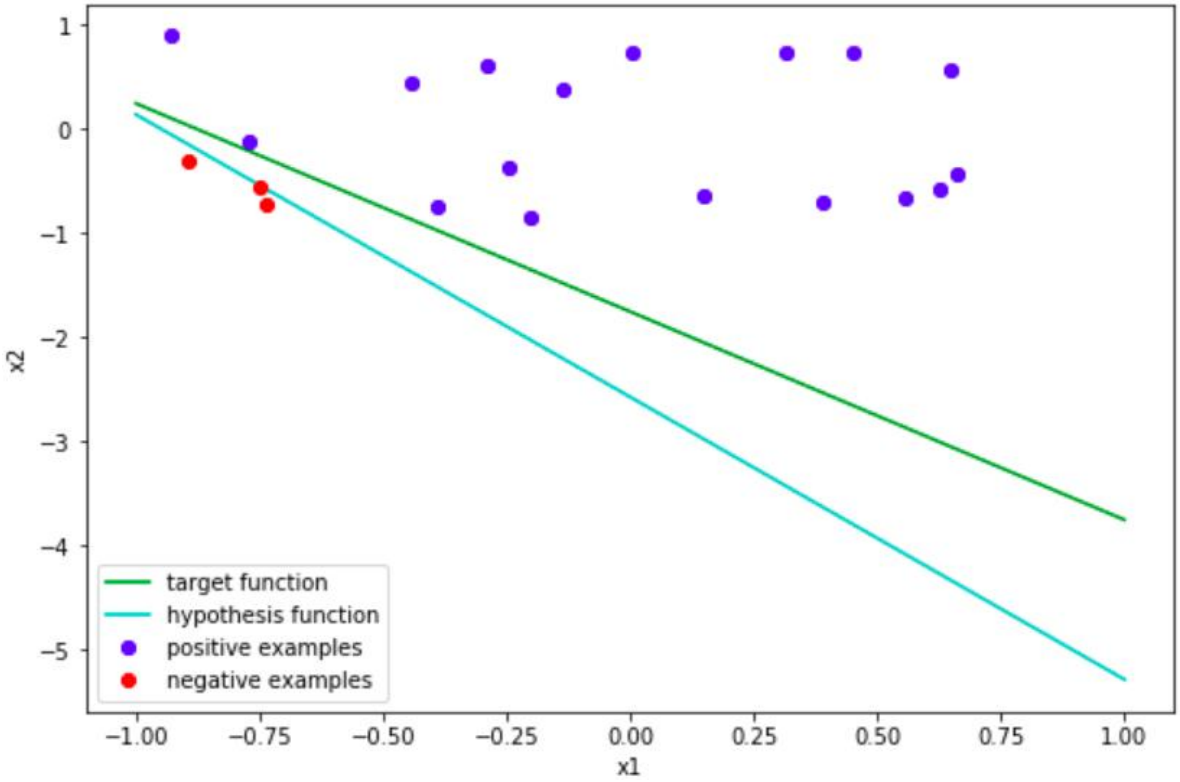

Figure 2: The graph for the second data set of size 20

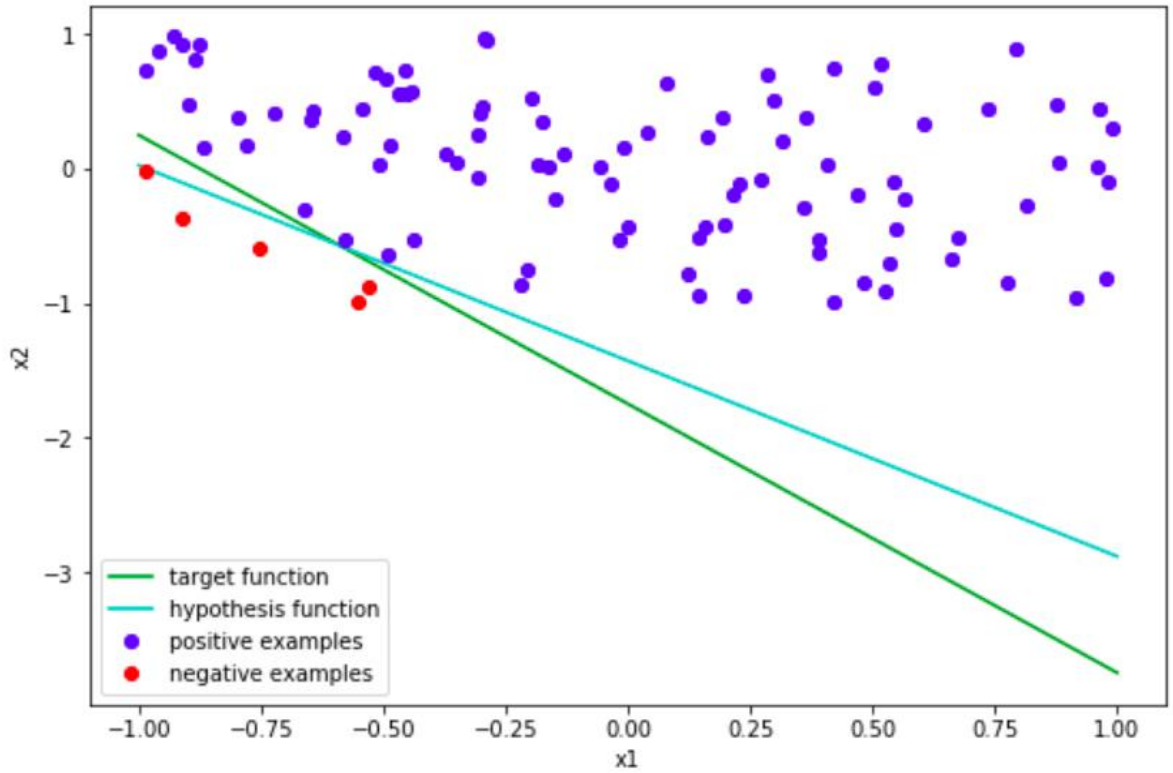

Figure 3: The graph for the data set of size 100

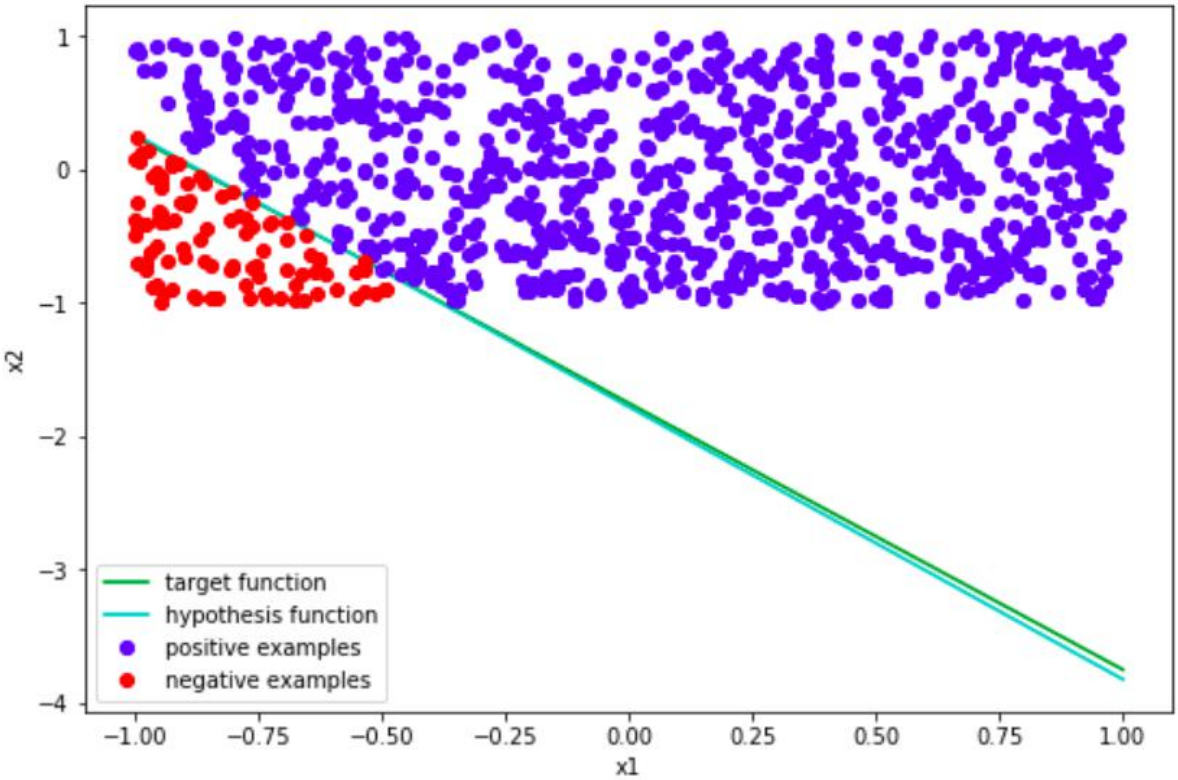

Figure 4: The graph for the data set of size 1000

If we compare the results, we can observe that the hypothesis is the closest to the target function when the data set is large enough.

B) Probability Related Problems

0.4 Independence

Intuitively, two random variables X and Y are "independent" if knowledge of the value of one tells you nothing at all about the value of the other. Precisely, if X and Y are discrete, independence means that P (X = x; Y = y) = P (X = x)P (Y = y), and if they are continuous, f (X = x; Y = y) = f (X = x)f (Y = y). Show the following, for independent random variable X and Y:

E[XY] = E[X]E[Y] for discrete and continuous cases respectively.

Solution:

Discrete Case:

$E[X] = \sum(x * P(X = x))$;

$E[Y] = \sum(y * P(Y = y))$;

$E[XY] = \sum\sum(x * y * P(X = x; Y = y))$;

Since X and Y are independent $P(X = x; Y = y) = P(X = x)*P(Y = y)$

$E[XY] = \sum\sum(x * y * P(X = x) * P(Y = y))$

$E[XY] = \sum\sum(x * P(X = x) * y * P(Y = y))$

$E[XY] = \sum(x * P(X = x)) * \sum y * (P(Y = y))$

$E[XY] = E[X] * E[Y]$

Continuous Case:

$E[X] = \int(x * f(X = x))dx$;

$E[Y] = \int(y * f(Y = y))dy$;

$E[XY] = \iint(x * y * f(X = x; Y = y))dxdy$

Since X and Y are independent $f(X = x; Y = y) = f(X = x)*f(Y = y)$

$E[XY] = \iint(x * y * f(X = x) * f(Y = y))dxdy$

$E[XY] = \iint(x * f(X = x) * y * f(Y = y))dxdy$

$E[XY] = \int(x * f(X = x))dx * \int y * (f(Y = y))dy$

$E[XY] = E[X] * E[Y]$

0.5 I.I.D. assumption in spam filters

Give at least 4 cases how "Independent and identically distributed random variables (i.i.d.)" assumption can be violated for spam filtering. Fore each case, please explain why i.i.d. is violated and how spammers might exploit the situation.

Answer:

1) Some of the words might appear more often and some rarer, therefore they are not equally distributed. Spammers can try to avoid the words that weight a lot.
2) There might be a dependency between the sender and the content. Spammers can send the spams from more reliable mail addresses.
3) Words might be dependent on each other. Spammers can try to predict the dependencies and avoid suspicious phrases.
4) Words with syntax errors might not be considered. Spammers can deliberately make syntax errors in some suspicious words.

0.6 *Spam filtering equation

Express Pr(S|W) in terms of the following components

1. Pr(S) is the overall probability that any given message is spam;

2. Pr(W|S) is the probability that the word "replica" appears in spam messages;

3. Pr(H) is the overall probability that any given message is not spam (is "ham");

4. Pr(W|H) is the probability that the word "replica" appears in ham messages.

Solution:

Pr(S|W) = Pr(S)*Pr(W) / Pr(W);

Pr(W) = Pr(W|S)*Pr(S) + Pr(W|H)*Pr(H);

Pr(W|S) = Pr(S)*Pr(W) / Pr(S) => Pr(S)*Pr(W) = Pr(W|S) * Pr(S);

Substitute Pr(S)*Pr(W) by Pr(W|S) * Pr(S), and Pr(W) by Pr(W|S)*Pr(S) + Pr(W|H)*Pr(H)

Answer:

Pr(S|W) = Pr(W|S)*Pr(S) / (Pr(W|S)*Pr(S) + Pr(W|H)*Pr(H))