Lab 3

Juan D Astudillo

11:59PM February 24, 2019

Perceptron

You will code the "perceptron learning algorithm". Take a look at the comments above the function. This is standard "Roxygen" format for documentation. Hopefully, we will get to packages at some point and we will go over this again. It is your job also to fill in this documentation.

```
#' Provide a name for this function
#' Perceptron Learning Algorithm
#' Explain what this function does in a few sentences
#' As a binary linear clasifier, Perseptrone will classify the given input data from a matrix. It will
#' @param Xinput      Matrix x with p number features n number of observations.
#' @param y_binary    Y is a binary vector of lenght n.
#' @param MAX_ITER    as default equals to 1000.
#' @param w           Weight is the initial numeric vector p+1.
#'
#' @return            The computed final parameter (weight) as a vector of length p + 1.
#' @export
#'
#' @author            Juan D Astudillo

perceptron_learning_algorithm = function(Xinput, y_binary, MAX_ITER = 1000, w = NULL){
p= ncol(Xinput)
n= nrow(Xinput)
if(is.null(w)){
   w = runif(ncol(Xinput))
}
for(iter in 1 : MAX_ITER){
  for(i in 1 : nrow(Xinput)){
    x_i = Xinput[i , ]
    yhat_i = ifelse(sum(x_i * w)>0 , 1, 0)
    w = w + as.numeric(y_binary[i] - yhat_i) * x_i
  }
}
w
}
```

To understand what the algorithm is doing - linear "discrimination" between two response categories, we can draw a picture. First let's make up some very simple training data $\mathbb{D}$.

```
Xy_simple = data.frame(
 response = factor(c(0, 0, 0, 1, 1, 1)), #nominal
 first_feature = c(1, 1, 2, 3, 3, 4),    #continuous
 second_feature = c(1, 2, 1, 3, 4, 3)    #continuous
)
```

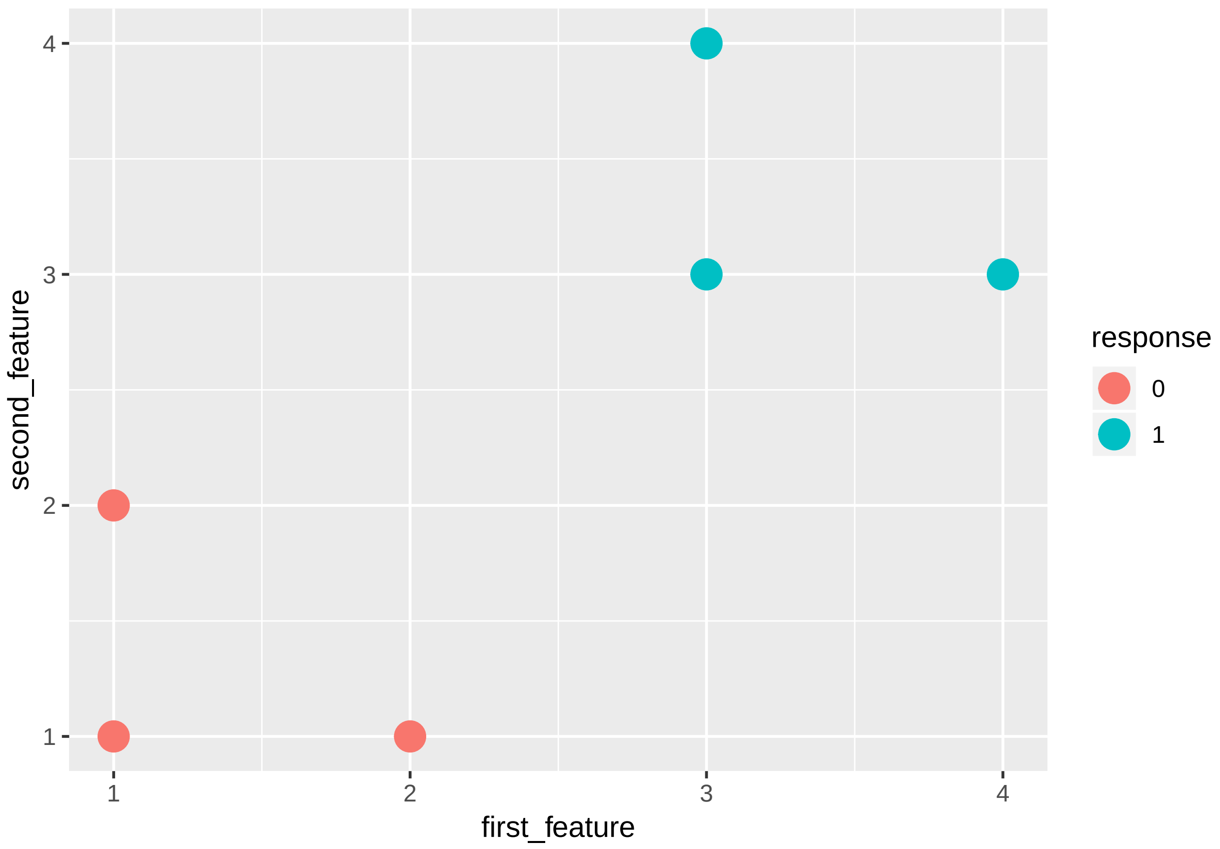
We haven't spoken about visualization yet, but it is important we do some of it now. First we load the visualization library we're going to use:

```
pacman::p_load(ggplot2)
```

We are going to just get some plots and not talk about the code to generate them as we will have a whole unit on visualization using `ggplot2` in the future.

Let's first plot y by the two features so the coordinate plane will be the two features and we use different colors to represent the third dimension, y.

```
simple_viz_obj = ggplot(Xy_simple, aes(x = first_feature, y = second_feature, color = response)) +
  geom_point(size = 5)
simple_viz_obj
```



Explain this picture. This shows the data that is linearly separable the green dots are the 1s and the red dots the 0s

Now, let us run the algorithm and see what happens:

```
w_vec_simple_per = perceptron_learning_algorithm(
  cbind(1, Xy_simple$first_feature, Xy_simple$second_feature),
  as.numeric(Xy_simple$response == 1))
w_vec_simple_per
```

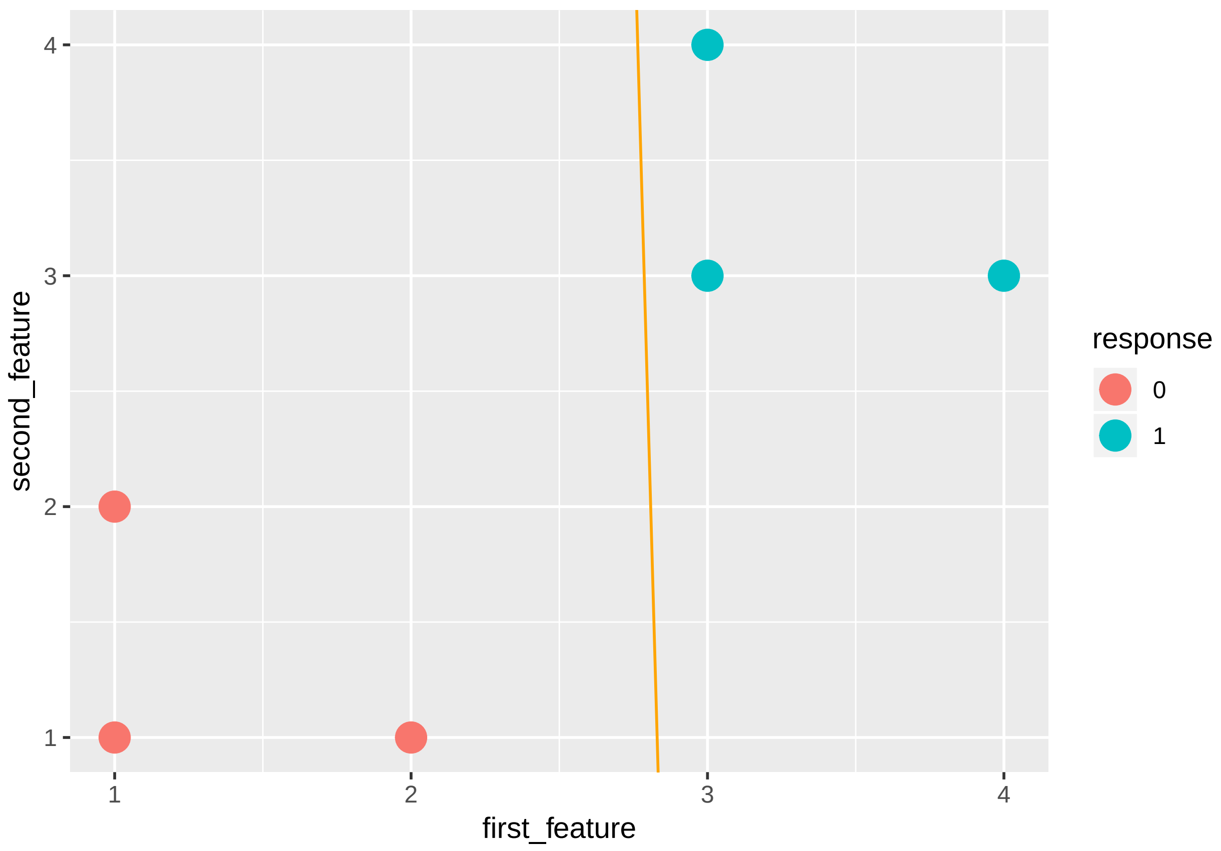
```
## [1] -8.41595203  2.95090208  0.06443916
```

TO-DO: Explain this output. What do the numbers mean? What is the intercept of this line and the slope? You will have to do some algebra.

these are the weights Wo= -10.589, W1= 3.155, W2= 1.003

(1.003)X2= 10.589 - (3.155)X1 X2= 10.56 - 3.15X1 the intersept is 10.56 and the slope -3.15

```
simple_perceptron_line = geom_abline(
    intercept = -w_vec_simple_per[1] / w_vec_simple_per[3],
    slope = -w_vec_simple_per[2] / w_vec_simple_per[3],
    color = "orange")
simple_viz_obj + simple_perceptron_line
```



Explain this picture. Why is this line of separation not "satisfying" to you? It is not satisfying.The line that we need is a line with the best separation, one that separates the 1s from the 0s. Here the line is over two data points doing a bad job. ## Support Vector Machine

```
X_simple_feature_matrix = as.matrix(Xy_simple[, 2 : 3])
y_binary = as.numeric(Xy_simple$response == 1)
```

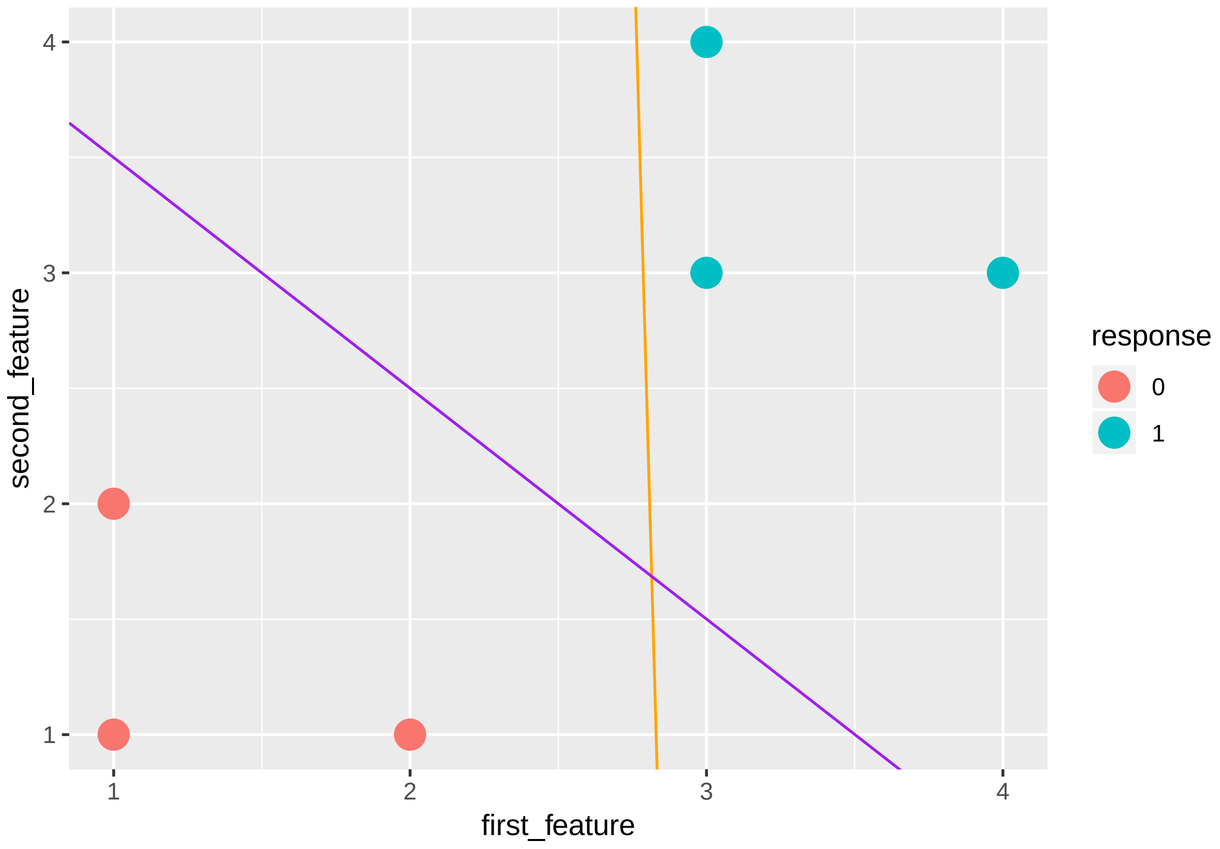
Use the `e1071` package to fit an SVM model to `y_binary` using the features in `X_simple_feature_matrix`. Do not specify the λ (i.e. do not specify the `cost` argument). Call the model object `svm_model`. Otherwise the remaining code won't work.

```
library(e1071)
lambda = 1e-9
n = nrow(X_simple_feature_matrix)
svm_model = svm(X_simple_feature_matrix, Xy_simple$response, kernel = "linear", cost = (2 * n * lambda)
```

and then use the following code to visualize the line in purple:

```
w_vec_simple_svm = c(
  svm_model$rho, #the b term
  -t(svm_model$coefs) %*% X_simple_feature_matrix[svm_model$index, ] # the other terms
)
```

```r
simple_svm_line = geom_abline(
    intercept = -w_vec_simple_svm[1] / w_vec_simple_svm[3],
    slope = -w_vec_simple_svm[2] / w_vec_simple_svm[3],
    color = "purple")
simple_viz_obj + simple_perceptron_line + simple_svm_line
```



Is this SVM line a better fit than the perceptron? Yes, the svm line is better than the perseptron line since it is in the middle of the separated data. The margin between the suppor vectors is bigger, the hyperplane is better located in the center.

3. Now write pseudocode for your own implementation of the linear support vector machine algorithm respecting the following spec making use of the nelder mead `optimx` function from lecture 5p. It turns out you do not need to load the package `neldermead` to use this function. You can feel free to define a function within this function if you wish.

Note there are differences between this spec and the perceptron learning algorithm spec in question #1. You should figure out a way to respect the `MAX_ITER` argument value.

```r
#' Support Vector Machine
#
#' This function implements the hinge-loss + maximum margin linear support vector machine algorithm of
#'
#' @param Xinput      The training data features as an n x p matrix.
#' @param y_binary    The training data responses as a vector of length n consisting of only 0's and 1'
#' @param MAX_ITER    The maximum number of iterations the algorithm performs. Defaults to 5000.
#' @param lambda      A scalar hyperparameter trading off margin of the hyperplane versus average hinge
#'                    The default value is 1.
#' @return            The computed final parameter (weight) as a vector of length p + 1
```

```
linear_svm_learning_algorithm = function(Xinput, y_binary, MAX_ITER = 5000, lambda = 0.1){
#

#
}
```

If you are enrolled in 390 the following is extra credit but if you're enrolled in 650, the following is required. Write the actual code. You may want to take a look at the `optimx` package we discussed in class.

```
#' This function implements the hinge-loss + maximum margin linear support vector machine algorithm of
#'
#' @param Xinput      The training data features as an n x p matrix.
#' @param y_binary    The training data responses as a vector of length n consisting of only 0's and 1'
#' @param MAX_ITER    The maximum number of iterations the algorithm performs. Defaults to 5000.
#' @param lambda      A scalar hyperparameter trading off margin of the hyperplane versus average hinge
#'                    The default value is 1.
#' @return            The computed final parameter (weight) as a vector of length p + 1
linear_svm_learning_algorithm = function(Xinput, y_binary, MAX_ITER = 5000, lambda = 0.1){
}
```

If you wrote code (the extra credit), run your function using the defaults and plot it in brown vis-a-vis the previous model's line:

```
svm_model_weights = linear_svm_learning_algorithm(X_simple_feature_matrix, y_binary)
my_svm_line = geom_abline(
    intercept = svm_model_weights[1] / svm_model_weights[3],#NOTE: negative sign removed from intercept
    slope = -svm_model_weights[2] / svm_model_weights[3],
    color = "brown")
```

```
## Error in -svm_model_weights[2]: invalid argument to unary operator
```

```
simple_viz_obj  + my_svm_line
```

```
## Error in eval(expr, envir, enclos): object 'my_svm_line' not found
```

Is this the same as what the `e1071` implementation returned? Why or why not?

4. Write a $k = 1$ nearest neighbor algorithm using the Euclidean distance function. Respect the spec below:

```
#' This function implements the nearest neighbor algorithm.
#'
#' @param Xinput      The training data features as an n x p matrix.
#' @param y_binary    The training data responses as a vector of length n consisting of only 0's and 1'
#' @param Xtest       The test data that the algorithm will predict on as a n* x p matrix.
#' @return            The predictions as a n* length vector.

nn_algorithm_predict = function(Xinput, y_binary, Xtest){
  prediction = c(rep(NA, nrow(Xtest)))
  i_star = c(rep(NA, nrow(Xtest)))
  for (k in 1:nrow(Xtest)){
    best_sqd_distance = Inf #good place to begin
    for (i in 1 : nrow(Xinput)){
      totdsqd = 0

      for (j in 1:ncol(Xinput)){
        dsqd = (Xinput[i, 1] - Xtest[k, 1])^2
        totdsqd = totdsqd + dsqd
      }
      if (dsqd < best_sqd_distance){
        best_sqd_distance = dsqd
        i_star[k] = i
      }
    }
    prediction[k] = y_binary[i_star[k]]
  }
  prediction
}
```

Write a few tests to ensure it actually works:

```
Xy = na.omit(MASS::biopsy) #The "breast cancer" data with all observations with missing values dropped
X = Xy[, 2 : 10] #V1, V2, ..., V9
y_binary = as.numeric(Xy$class == "malignant")

test = matrix(c(7.8,5.2),2,1)
nn_algorithm_predict(X,y_binary, test)
```

```
## [1] 1 0
```

We now add an argument `d` representing any legal distance function to the `nn_algorithm_predict` function. Update the implementation so it performs NN using that distance function. Set the default function to be the Euclidean distance in the original function. Also, alter the documentation in the appropriate places.

```
nn_algorithm_predict = function(Xinput, y_binary, Xtest, d = euclid){
  euclid= function(x1, x2){
    ((x1 - x2)^2)
  }

  prediction = c(rep(NA, nrow(Xtest)))
  i_star = c(rep(NA, nrow(Xtest)))
  for (k in 1:nrow(Xtest)){
    best_sqd_distance = Inf #good place to begin
    for (i in 1 : nrow(Xinput)){
      totdsqd = 0

      for (j in 1:ncol(Xinput)){
        dsqd = euclid(Xinput[i, 1],Xtest[k,1])
        totdsqd = totdsqd + dsqd
      }
      if (dsqd < best_sqd_distance){
        best_sqd_distance = dsqd
        i_star[k] = i
      }
    }
    prediction[k] = y_binary[i_star[k]]
  }
  prediction
}
```

For extra credit (unless you're a masters student), add an argument `k` to the `nn_algorithm_predict` function and update the implementation so it performs KNN. In the case of a tie, choose $\hat{y}$ randomly. Set the default `k` to be the square root of the size of $\mathcal{D}$ which is an empirical rule-of-thumb popularized by the "Pattern Classification" book by Duda, Hart and Stork (2007). Also, alter the documentation in the appropriate places.

```
#TO-DO --- extra credit for undergrads
```