assignment_03_solution

March 4, 2022

Please fill in your name and that of your teammate.

You:

Teammate:

1 Introduction

Welcome to the third lab. There is much to go through today so we will keep extra concepts to a minimum. There is no new library introduced at this lecture as we will keep using `numpy` for the heavy lifting, `scikit-learn` for the algorithms, and `seaborn` / `matplotlib` for plotting. Careful about reusing variable names in the notebook and computing cells out of order: frequent calls to `Kernel -> Restart and Run All` can save you from headaches.

The assignment starts getting math-heavy. Here's a new tool to aid you with the debugging. Explicitly `import IPython` at the beginning of a notebook (or Python file) to have access to the computational Python kernel. You can then call `IPython.embed()` at an arbitrary place in your code (say, inside a loop) and it will pause the computation and drop you into an interactive console. You can then evaluate Python code in the context where it was called. Here is an example:

```
import numpy as np
for i in range(10):
    guess = np.random.normal()
    function_that_fails_because_of(guess, i)
```

Let's say your function fails for `i==9`, how would you find the error? Typically you may want to edit the code and print `guess` and `i` to see what is happening, but it is slow and passive. What if you want to try to pass `i+1` and see if that works? What if you want to try a few other random numbers with the same `i`? Enter `IPython.embed()`:

```
import numpy as np
import IPython
for i in range(10):
    guess = np.random.normal()
    if i==9: IPython.embed()
    function_that_fails_because_of(guess, i)
```

If you execute this code, the cell output will show an interactive console in your output cell. Here you can send commands to be interpreted by the Python kernel of the notebook. You could then try something like the following lines for example (one at a time):

```
i #=> prints value of i
guess  #=> prints value of guess
function_that_fails_because_of(guess, i) #=> fails and shows you the error
function_that_fails_because_of(guess, i+1) #=> change parameters: will it work?
function_that_fails_because_of(guess, i-1) #=> what about this one?
guess = np.random.normal() #=> overwrites the value of `guess` in the kernel
function_that_fails_because_of(guess, i) #=> will this work this time?
```

As you can see you can test your code in the context of the function (or, here, loop), find the code that works, then you can go ahead and copy+paste in your actual code. If you need to exit the console and resume the computation (with whatever change you executed, as the kernel is the same) just type `exit()`.

Bonus: you can ask the kernel to drop you into a *debugger* session every time you get an error. This can be tricky with Jupyter Notebooks, so use with caution, but can also be a lifesaver if you are willing to learn about postmortem debugging. You do that by adding the following lines on top of your code (need to execute them only once):

```
# DEBUG: uncaught exceptions drop you into ipdb for postmortem debugging
import sys, IPython; sys.excepthook = IPython.core.ultratb.ColorTB(call_pdb=True)
```

After this line, if you encounter an error or uncaught exception in your code, rather than terminating you will be dropped in an `ipdb` (fancier version of `pdb`) console where you can interrogate the program about the conditions causing the crash.

Good hunting!

1.0.1 How to pass the lab?

Below you find the exercise questions. Each question awarding points is numbered and states the number of points like this: **[0pt]**. To answer a question, fill the cell below with your answer (markdown for text, code for implementation). Incorrect or incomplete answers are in principle worth 0 points: to assign partial reward is only up to teacher discretion. Over-complete answers do not award extra points (though they are appreciated and will be kept under consideration). Save your work frequently! (`ctrl+s`)

You need at least 22 points (out of 33 available) to pass (66%).

2 1. Fundamentals

Let's make sure some of the core points are clear before addressing the specific algorithms.

1.1 [1pt] Write the equation of the Gaussian density. Use Latex inside the Markdown cell. I suggest you type it out rather than copy+paste from the Internet: the goal of this question is to *force* you to read one term at a time, and understand which is clear to you and which is not. For example, the equation you will write here a norm, while our later applications of the formula do not, since we will actually be using the Gaussian Probability Density Function (PDF) equation instead. Do you understand why? Did you study this before?

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$$

1.2 [1pt] Explain why we maximize the log-likelihood rather than the likelihood. In particular, what is the advantage in using the log **rather than another operation?** We use the log because the likelihood has a product, which is hard to maximize, while the log transforms it into a much easier sum. The choice for the logarithmic is because the dependent term of the Gaussian is an exponential.

1.3 [1pt] Why the equation maximizing the log-likelihood of a Gaussian does not include the parameter σ? Because the terms with σ do not depend on w and thus do not influence the $argmax$.

1.4 [1pt] Explain the meaning of i.i.d. (in English), using the simplest words you can. Identically and independently distributed means that random values drawn from such a distribution do not influence each other. Drawing one does not change the probability of drawing the next one.

1.5 [2pt] Write the equation of the Bayes' Rule (use Latex). Then write below how to read it in English.

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

The probability of class y given the input x is equal to: the probability of the class y generating such an input x, multiplied by the prior probability of selecting class y among all classes, divided by the overall probability of having input x.

3 2. Linear Regression

2.1 [2pt] Explain the meaning of $y_i = \langle w, x_i \rangle + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$ (in English). Utilize the word 'prototype'. The label y_i of an input x_i can be interpreted as class prototype $\langle w, x_i \rangle$ plus a certain error ϵ drawn from a Gaussian (or normal) distribution centered in zero $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The inverse is also true, as obvious by swapping the side of ϵ in the equation: the prototype predicted by the model with $\langle w, x_i \rangle$ corresponds to the exact label y_i plus some normal noise ϵ (which is centered in zero and symmetric, so the $+$ or $-$ sign does not matter).

Enough theory, let's get our hands in there. Since we are working with a regression task, let us generate some data from an underlying linear function with some noise.
[think: the process below is (correct but) unnecessary convoluted: would you be able (yet) to simplify it? Always prefer simpler code!]

[1]:
```
import numpy as np
trg_fn = lambda x: 2*x - 1 # hi I'm lambda, remember me?
some_noise = lambda: np.random.normal(0,2) # Gaussian noise with mu=0, sigma=2
# Below we add a 1 to every row as the bias (constant) input
# Think about each part and discuss if you do not understand something yet!
data = np.array([[x, 1, trg_fn(x) + some_noise()] for x in np.linspace(-10, 10,␣
↪50)])
```

```
*x, y = data.transpose() # easier using splat and numpy: do you understand it?
x = np.array(x).transpose() # back to *rows* with input features and bias input
```

2.2 [1pt] Write a (Python) function that takes a data point in input and returns the squared error Loss. Test it by using a constant prediction model $y = 1$ and compute the Risk over all the data.

[2]:
```
def sq_loss(trg, pred): return (trg - pred)**2
sq_loss_risk = sum(sq_loss(trg, 1) for trg in y)
print(f"Risk for constant predictor 'y=1' using squared errors loss:␣
 ↪{sq_loss_risk}")
```

```
Risk for constant predictor 'y=1' using squared errors loss: 8367.948095222817
```

- Numpy's linear algebra library provides matrix inversion, but you should use instead the pseudo-inverse to cope with singular covariance matrices: `np.linalg.pinv()`
- Numpy's array provides inner product with the function `dot()`
- Typically `dot()` will find the right direction for one-dimensional arrays, which means that you should never need to transpose them
- Using `dot()` with matrices instead *always* requires you to `transpose()` to the right orientation! Write the math and keep track of what you are doing.
- Remember that matrix product is not commutative: `A.dot(B)` is NOT equal to `B.dot(A)`. Refresh also how `A.dot(B)` requires the number of columns of `A` to be the same as the number of rows of `B`, and the result will have the same number of rows of `A` and the number of columns of `B`.
- Linear regression uses a closed-form solution, **not a loop**, and you will not use the implementation of the loss above in the rest of the assignment (because it is implicit in the algorithm's solution).

2.3 [2pt] Write a function that takes in input a list of data points and a list of labels, and returns the w vector using the closed-form solution of Linear Regression. Test it on the data above and print the computed w.

[3]:
```
def lr_w(x, y): return np.linalg.pinv(x.transpose().dot(x)).dot(x.transpose()).
 ↪dot(y)
w = lr_w(x, y)
print(w)
```

```
[ 2.13755392 -0.86214539]
```

2.4 [2pt] Predict the labels for all points using your Linear Regression implementation and the `w` vector from the previous question. For each data point, print the triplet of `(label, prediction, loss)`. Then compute and print the (squared error) Risk over the dataset. Remember that you need to build the linear model (careful handling the bias), then you are doing *regression* not *classification*, which means you are predicting the value $\hat{y}$ based on your x. You should get back m and q close to 2 and -1, and the risk roughly proportional to the square of the expected error multiplied by the number of points (or close below that).
[think: what is the expected error? Do you understand why? You can generate the points and compute the risk multiple times to verify your hypothesis]

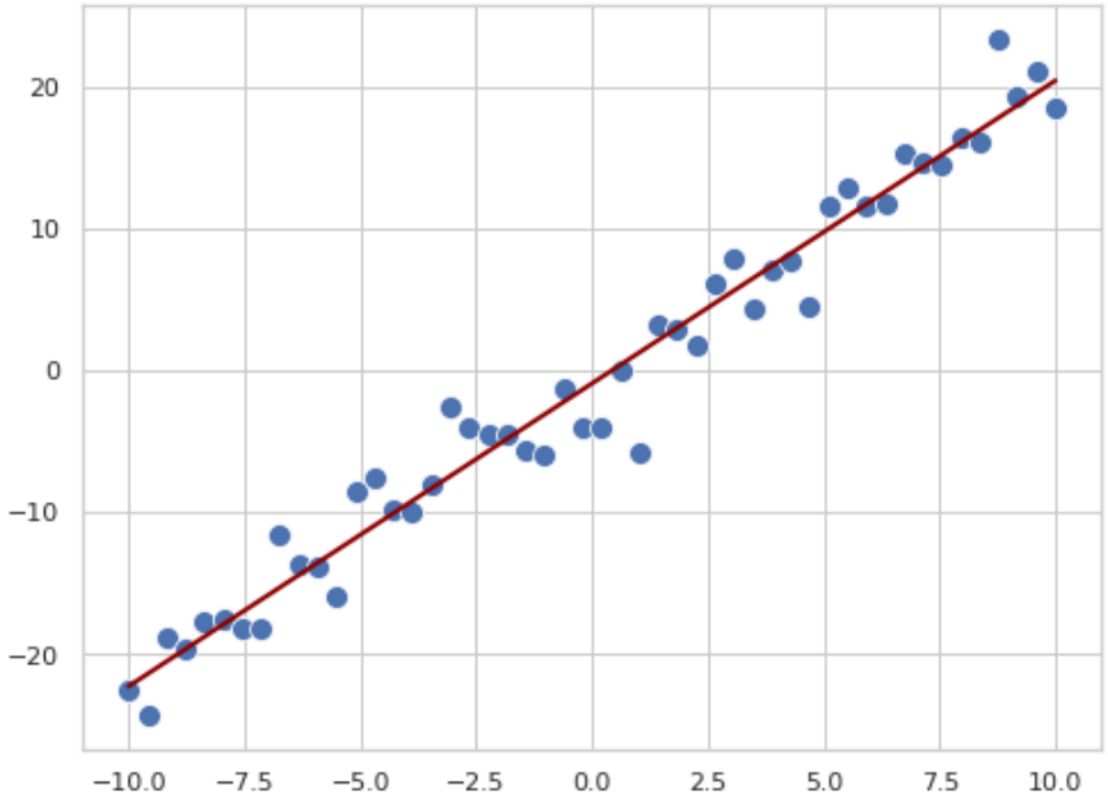
[4]:
```
predictions = list(map(round, w.dot(x.transpose())))
pred_pairs = list(zip(y, predictions))
losses = [round(sq_loss(pr, tr), 2) for tr, pr in pred_pairs]
triplets = list(zip(y, predictions, losses))
print(triplets)
risk = sum(losses)
print(f"\nRisk for trained predictor 'y = {round(w[0], 2)}x +"
      f"({round(w[1], 2)})' using squared errors loss: {round(risk, 2)}")
```

```
[(-22.453864154607444, -22, 0.21), (-24.355288554468324, -21, 11.26),
(-18.830239122972035, -20, 1.37), (-19.665900159366668, -20, 0.11),
(-17.590594067916417, -19, 1.99), (-17.48860480323548, -18, 0.26),
(-18.178304958883956, -17, 1.39), (-18.089088659308935, -16, 4.36),
(-11.52903912270708, -15, 12.05), (-13.717395400343177, -14, 0.08),
(-13.844148118394871, -14, 0.02), (-15.831761950240315, -13, 8.02),
(-8.526387229737626, -12, 12.07), (-7.461847746131552, -11, 12.52),
(-9.715655854404257, -10, 0.08), (-9.909317551003593, -9, 0.83),
(-8.045928551437639, -8, 0.0), (-2.531812785071322, -7, 19.96),
(-3.9803803043975448, -7, 9.12), (-4.47962380808473, -6, 2.31),
(-4.47374770595854, -5, 0.28), (-5.640332273994828, -4, 2.69),
(-5.9630644437187845, -3, 8.78), (-1.2021304001848205, -2, 0.64),
(-4.0095226892542115, -1, 9.06), (-3.946673301914443, 0, 15.58),
(0.0499153783108017, 0, 0.0), (-5.7507453206489565, 1, 45.57),
(3.309397567346861, 2, 1.71), (2.908114618478713, 3, 0.01), (1.8155796455017206,
4, 4.77), (6.158029201212644, 5, 1.34), (7.920590973028711, 6, 3.69),
(4.423001850855109, 7, 6.64), (7.074166622769325, 7, 0.01), (7.86603442423755,
8, 0.02), (4.536624663074363, 9, 19.92), (11.587999224641994, 10, 2.52),
(12.944521061201936, 11, 3.78), (11.71775175635889, 12, 0.08),
(11.84221548524487, 13, 1.34), (15.430996342049463, 14, 2.05),
(14.662467467973228, 14, 0.44), (14.533052724343113, 15, 0.22),
(16.422705771229143, 16, 0.18), (16.140145639855614, 17, 0.74),
(23.436430058554016, 18, 29.55), (19.470424677222745, 19, 0.22),
(21.204989635849472, 20, 1.45), (18.64897471130482, 21, 5.53)]

Risk for trained predictor 'y = 2.14x +(-0.86)' using squared errors loss:
266.82
```

2.5 [2pt] Plot the data and the model. You should be able to partially reuse the printing code from the last lab (particularly plot and params-to-boundary conversion), but feel free to customize it as you need. I will keep repeating this for a while more: careful with the bias!
(Think: you can plot the model's predictions very easily if you use linear algebra, do you understand what is `x.dot(w)`?)

[5]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(rc={'figure.figsize':(8,6)}, style="whitegrid")
```

```
# start a variable name with _ to declare you do not intend to use it (_ itself␣
 ↪is a valid name)
actual_x, _bias = x.transpose()

# did you learn to write your own plotting methods?
def lr_data_plot(): sns.scatterplot(x=actual_x, y=y, s=100)
def lr_model_plot(w, color='darkred', linewidth=2):
    # Seaborn utilized matplotlib underneath. You can grab `ax` objects for␣
 ↪further syling.
    sns.lineplot(x=actual_x, y=x.dot(w), color=color, linewidth=linewidth)

lr_data_plot()
lr_model_plot(w)
```



2.6 [2pt] Find Linear Regression in scikit-learn and train a model on the data. The input to the `fit()` function should be a matrix and a vector, so try forcing `actual_x` into a $n \times 1$ matrix by using `actual_x.reshape((-1, 1))`. If you want to predict the outputs, you simply need to pass the same data matrix to the method `predict()`.

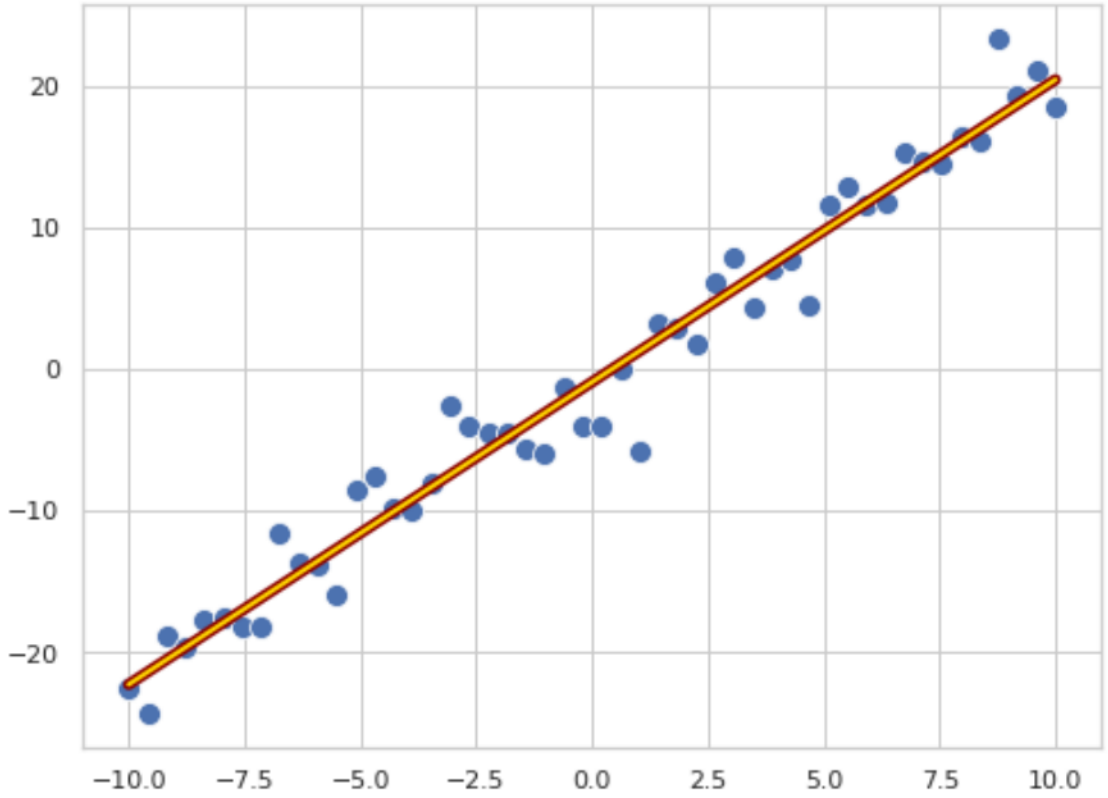
[6]:

```
from sklearn.linear_model import LinearRegression
x_mat = actual_x.reshape((-1, 1)) # do you understand this? can you change it␣
 ↪for the exam?
trained = LinearRegression().fit(x_mat, y)
preds = trained.predict(x_mat)
```

2.7 [1pt] Plot in a single figure: (i) the data points as a scatterplot; (ii) the model you learned using your implementation of Linear Regression; (iii) the model you trained using the scikit-implementation. Careful, it may be that you plot both but they are exactly the same, so they are superimposed and you only see one line. To verify this you can try plotting them with different colors and different thickness, or changing the linestyle (e.g. one normal line and one dashed line). A list of available linestyles can be found in the matplotlib documentation.

To get the parametrization of the sklearn implementation use the following:
`w_skl = [trained.coef_[0], trained.intercept_]`
Do you understand what this does? Can you see how this corresponds to what we did last week? In future assignments you will need to find the model parametrization by yourself!

[7]:
```
lr_data_plot()
lr_model_plot(w, linewidth=5)
w_skl = [trained.coef_[0], trained.intercept_]
lr_model_plot(w_skl, color='gold', linewidth=2)
```



4 3. Linear Discriminant Analysis

We switch now into *binary classification*. Let's load the iris dataset once again for this exercise, and carefully selecting the data to have an easy binary classification problem (for now). Notice we are **not** using classes $-1, +1$ anymore, because we do not need to compute a Margin here.

[8]:
```
import numpy as np
from sklearn.datasets import load_iris
iris_x, iris_y = load_iris(return_X_y=True) # print these points to understand␣
 ↪them!
x1 = np.array([r[0] for r in iris_x]) # first feature
x2 = np.array([r[2] for r in iris_x]) # third feature

x = np.array([x1, x2]).transpose() # numpy gives us transpose() for free
# Reduce the three classes into two for binary classification {1, 2}
y = np.array([1 if y in [1,2] else 2 for y in iris_y])
```

To solve LDA we need to find the parametrization $\theta_y = (\mu_y, \Sigma, \pi_y)$. Since μ and π are class-dependent, remember to first split the input data based on which class it belongs to.

3.1 [1pt] Write a (Python) function that takes a dataset (inputs and labels) in input and returns a dictionary hashing each of the m classes to the a list of the points belonging to that class. We will call this `partition` in the next questions. Hint: the method `dict.get(<key>, <def>)` can be used to fetch values from a dictionary same as `dict[<key>]`, but when the key is not found it returns the second argument, which is the *default value*... what if you pass an empty list...

[9]:
```
def partition_dset(inputs, classes):
    ret = {}
    for inp, cls in zip(inputs, classes):
        ret[cls] = ret.get(cls, []) + [inp]
    return ret

part = partition_dset(x, y)
```

3.2 [1pt] Write a function that takes the `partition` in input and returns a dictionary hashing each class to its corresponding prototype μ_y. Hint: function `dict.items()` returns a list of pairs `(key, value)` from the dictionary. Then a *dictionary comprehension* works same as a list comprehension, with a `for` loop that generates elements. Only this time you need to pass both a key and a value like this:

```
d = { the_key: compute_value(a, b) for a, b in another_dict.items() }
```

If you can use that then the answer is basically one line. If it is complicated instead, write explicit `for` loops, which are exactly equivalent (and perhaps even more readable:

```
d = {}
for a, b in another_dict.items():
    d[the_key] = compute_value(a, b)
```

Just find a style that is comfortable with you. May take a few weeks but you will get there.

[10]:
```
def lda_mus(part): return { cls: sum(pts)/len(pts) for cls, pts in part.items()␣
 ↪}
mus = lda_mus(part)
```

3.3 [1pt] Write a function that takes the `partition` in input and returns a dictionary hashing each class to its corresponding prior π_y.

[11]:
```
def lda_pis(part):
    tot_len = sum(map(len, part.values()))
    return {cls: len(pts)/tot_len for cls, pts in part.items()}
pis = lda_pis(part)
```

3.4 [4pt] Write a function that takes the `partition` in input and the class-wise center estimates (the means from above) and returns the corresponding Σ (one for all classes and all inputs). You may need to use `np.concatenate()` to join the $x_i - \hat{\mu}_{y_i}$ from each class. Print the `array.shape` (property not method so no `()`) to verify if your linear algebra is on point so far: the covariance matrix between all inputs should have as many rows (and columns) as the number of features (hint: that's 2). Example: `assert sigma.shape == (2,2)` should not raise an error.

[12]:
```
def lda_sigma(part, mus):
    x_minus_mu = np.concatenate([pts-mus[cls] for cls, pts in part.items()])
    return x_minus_mu.transpose().dot(x_minus_mu) / len(x_minus_mu)
sigma = lda_sigma(part, mus)
assert sigma.shape == (2,2)
```

3.5 [5pt] Write the function for the decision boundary of LDA $f(x)$. You can find the logarithm in the `math` module: `from math import log` then just `l = log(a)`. You need to implement the equations for w and q from the slides: this cell will be very math heavy, be careful though and you should be able to get it right in few lines. Remember to check for the shape of w and b to verify if your matrix products are computed the right way. You want w to be of length 2, and b is a scalar.

[13]:
```
from math import log
sigma_inv = np.linalg.pinv(sigma)
lda_w = sigma_inv.dot(mus[1] - mus[2])
lda_b = 0.5*mus[2].dot(sigma_inv).dot(mus[2]) - \
        0.5*mus[1].dot(sigma_inv).dot(mus[1]) + \
        log(pis[1]) - log(pis[2])
print(f"w: {lda_w}\nb: {lda_b}")
f = lambda x: lda_w.dot(x) + lda_b
```

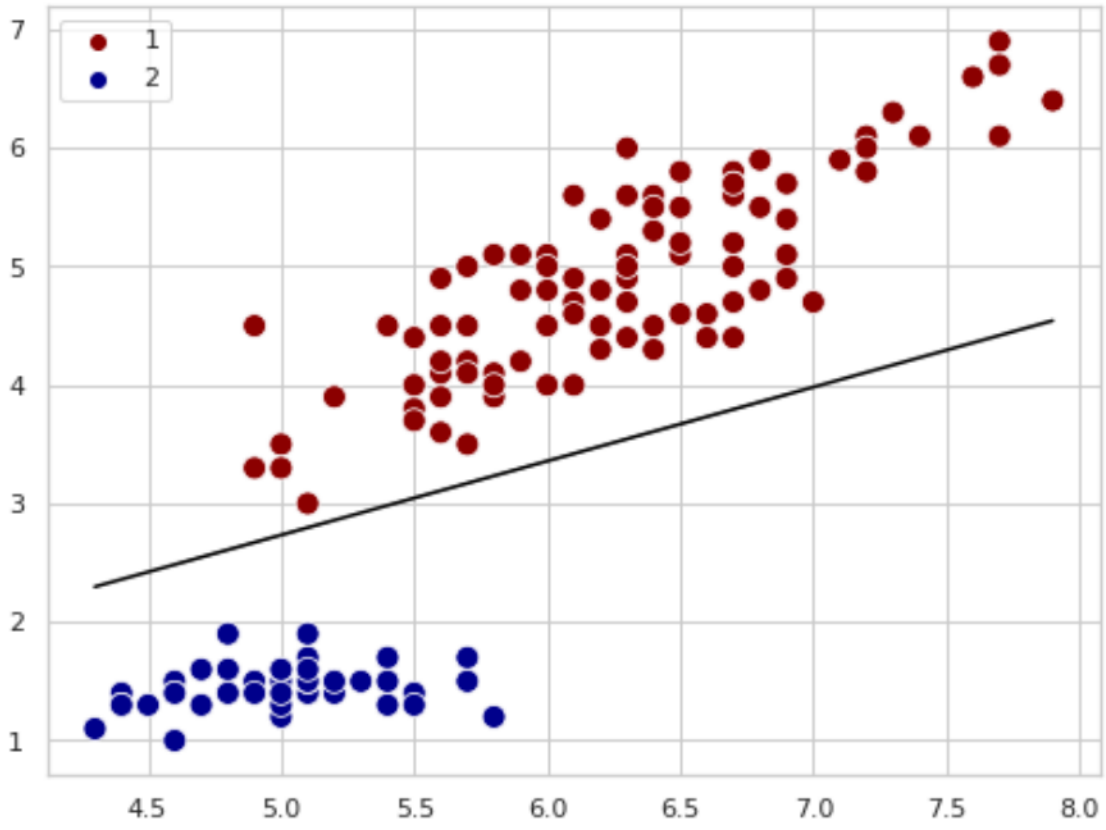
```
w: [-7.96528842 12.76916894]
b: 4.912548207950932
```

3.6 [2pt] Plot the LDA decision boundary on top of the data.

- Notice that we are using different data from the Linear Regression questions above, and that our model now generated w and b separately, so you **need** to adapt the plotting functions – actually it's quicker to just rewrite them.
- Plotting is one of few applications where generalizing your code just makes for a stunted replica of the original (better) interface, so specialize when needed but for different applications just write a new one rather than reusing your code.
- Remember that we already had a good plotting function for a classification problem just last week, why don't you check it out again?
- Finally, remember that you already have code that converts a parametrization from w and b to m and q (last assignment), you can simply copy+paste it here to simplify your generation of the model's points.

[14]:
```python
def wb2mq(w, b):
    assert len(w) == 2, "This implementation only works in 2D"
    assert w[0] != 0 and w[1] != 0 and b != 0 # simplify
    return [w[0]/-w[1], b/-w[1]] # m and q

def params2boundary(w, b):
    m, q = wb2mq(w, b)
    print(f"m: {m}, q: {q}")
    return lambda x: m*x + q

def lda_data_plot():
    sns.scatterplot(x=x1, y=x2,
      hue=y, # let's use different colors for the two classes
      palette=sns.color_palette(['darkred', 'darkblue']),
      s=100)

def lda_model_plot(w, b, color='black'):
    model = params2boundary(w, b)
    model_points = [model(inp) for inp in x1]
    sns.lineplot(x=x1, y=model_points, color=color)

lda_data_plot()
lda_model_plot(lda_w, lda_b)
```

```
m: 0.6237906674023641, q: -0.3847194935007181
```

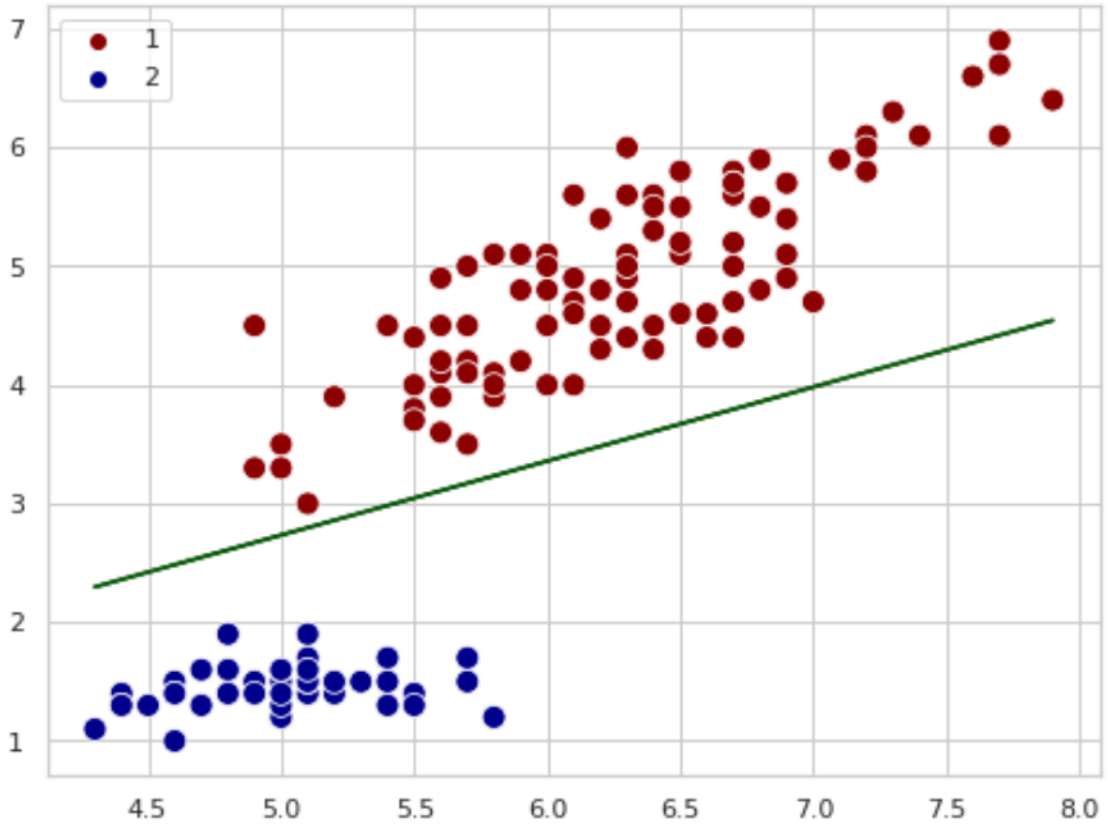


3.7 [1pt] Find LDA on scikit-learn; train a model on the data and add it to the print above (data + model from your implementation). You ned to pass the correct `solver` parameter to the sklearn constructor, check the documentation to understand what I mean. If you do not the result should still look exactly the same as your implementation (because the data is linearly separable), but you should be aware of which technique your library uses and we have not gotten to SVD yet.
Remember to make sure that you can distinguish the two boundaries even if they overlap (e.g. use different colors). If you use our conversions from last exercise you should also see the printed values of m and q and they are likely to differ in the least significant digits even though the graph looks the same.
Also consider that LDA is a **multiclass method**, and so its parametrization is in principle a list for the many boundaries: you need to access the coefficients of the *first* (and here, only) boundary using `[trained_model.coef_[0], trained_model.intercept_[0]]`.

[15]:
```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
print(f"x shape: {x.shape} y shape: {y.shape}")
trained = LinearDiscriminantAnalysis('lsqr').fit(x, y)
w_skl, b_skl = [trained.coef_[0], trained.intercept_[0]]
lda_data_plot()
lda_model_plot(lda_w, lda_b, color='black')
```

```
lda_model_plot(w_skl, b_skl, color='darkgreen')
```

```
x shape: (150, 2) y shape: (150,)
m: 0.6237906674023641, q: -0.3847194935007181
m: 0.623790667402364, q: -0.38471949350071727
```



4.1 At the end of the exercise

Bonus question reward no points! Answering this will have no influence on your scoring, not at the assignment and not towards the exam score. But solving it will reward you with skills that will make the next lectures easier, give you real applications, and will be very good practice towards the exam.

The solution for this questions will not be included in the regular lab solutions pdf, but you are welcome to open a discussion on the Moodle: we will support your addressing it, and you may meet other students that choose to solve this, and find a teammate for the next assignment that is willing to do things for fun and not only for score :)

Let's see some multiclass classification. Copy the code loading the Iris dataset, you want to extract the same features (so you can plot in 2D), but keep the three classes.
[think: does it matter what label does each class have? Could you use strings such as `['a', 'b', 'c']`*?]*

Then run the scikit-learn LDA on the data to obtain a trained model. At this point you can open up the trained coefficients again, and rather than taking only the first like you did with `[trained.coef_[0], trained.intercept_[0]]`, you should have TWO w vectors and TWO b constants **per each pair** of classes. *[think: the space is actually split in several subspaces. Can you derive how many? Can you design a decision tree on top of the boundaries to do the classification as the number of boundaries grow?]*

BONUS [ZERO pt] Plot the boundaries classifying the three species of Iris in the dataset based on the two features used so far.

4.1.1 Final considerations

Stop for a moment and think how hard it was to derive these equations (in the lecture), and how hard it was instead to implement them (once you get them right). These are two very different skills.
To understand the derivation you need to think hard, express your concept in math (actually requiring broad knowledge of many of its subfields), see it through with absolute precision, and finally correctly solve the equations.
To implement the method, you need to map the math to the correct function calls (hard, but arguably less), and you only ever work with the final solution, but you deal with programming languages and libraries and documentations.
This is the reason why so many people nowadays broadly advertise machine learning skills after taking short tutorials. But if you do not understand what a parameter is for, you will only be guessing which value to use.

The reason why you are sweating so much on this course is to gain an edge over all of those who only ever learn to *use* the tools: by instead *making* the tools you understand them from the inside out, their applications and limitations, and even become capable of adapting and improving them. Keep up: this course is not easy, but machine learning has become unavoidable in your field, and these foundations will enable you to bend the whole field to your needs.