Rapport DM 1

Grégoire DHIMOÏLA

April 2023

Abstract

This document contains the report for the second homework of the Machine Learning course. It contains the answers to the questions and links to the code used to answer them. It also contains the main results of the experiments.

1 Question 1

1.1

It is important to regularize as it helps prevent overfitting and stupidly high weights which would be bad for generalization. It also helps to prevent the model to fit the noise.

1.2

/

1.3

The best value for λ would be before the capabilities start to degrade, and after the model norm started to shrink.

To tune it we can focus on the useful range of lambda and make a K-fold cross validation.

After making that, we find $\lambda = 0.1$ for the L1 regularization, and $\lambda = 5$ for L2.

1.4

The model is plotted for 4 values of lambda : before having any influence, when it is optimal, when the norm is lowest (and capabilities are extremely poor) and when it just diverged.

As expected, with weak lambda, the model is close to the linear regression model. As lambda increases, the model becomes more and more sparse. When lambda is optimal, we can clearly see bits of the images that are the most important for the classification, all the rest being set to 0. When lambda is too high, the model is too sparse and the accuracy decreases. Eventually, the model just diverges and becomes noise.

At optimality, the L1 regularisation was the strongest in setting as many coefficients as possible to 0. The L2 regularisation was more gentle and simply sort of smoothed out the image, which should be more desirable in a real world application.

I believe when lambda is too high, the learning rate is not small enough and the gradient steps make the model diverge, which is why we see noise at some point.

2 Question 2

2.1

Let $\pi = P(Y = 1)$.

En utilisant la formule de Bayes et avec $\pi = P(Y = 1)$:

$$\begin{aligned} P(Y = 1|X) &= \frac{P(X|Y = 1)\pi}{P(X)} \\ &= \frac{P(X|Y = 1)\pi}{P(X|Y = 1)\pi + P(X|Y = -1)(1 - \pi)} \\ &= \frac{1}{1 + \frac{P(X|Y=-1)}{P(X|Y=1)} \frac{1-\pi}{\pi}} \end{aligned}$$

Or on a :

$$\begin{aligned} \log\left(\frac{P(X|Y = 1)}{P(X|Y = -1)}\right) &= \frac{1}{2}\big((X - \mu_1)^T\Sigma^{-1}(X - \mu_1) - (X - \mu_{-1})^T\Sigma^{-1}(X - \mu_{-1})\big) \\ &= \frac{1}{2}\big(X^T\Sigma^{-1}(\mu_{-1} - \mu_1) - \mu_1^T\Sigma^{-1}(X - \mu_1) + \mu_{-1}^T\Sigma^{-1}(X - \mu_{-1})\big) \\ &= \frac{1}{2}\big(X^T\Sigma^{-1}(\mu_{-1} - \mu_1) + (\mu_{-1} - \mu_1)^T\Sigma^{-1}X + \mu_1^T\Sigma^{-1}\mu_1 - \mu_{-1}^T\Sigma^{-1}\mu_{-1}\big) \end{aligned}$$

On utilise le fait que $X^T\Sigma^{-1}(\mu_{-1} - \mu_1)$ est égal à sa transposé (car c'est un scalaire):

$$= \frac{1}{2}\big(- \langle\alpha_1, X\rangle - \alpha_0\big)$$

Avec:

$$\begin{aligned} \alpha_1 &= (\Sigma^{-1} + \Sigma^{-1^T})(\mu_1 - \mu_{-1}) \\ \alpha_0 &= \mu_1^T\Sigma^{-1}\mu_1 - \mu_{-1}^T\Sigma^{-1}\mu_{-1} \end{aligned}$$

On obtient finalement :

$$\begin{aligned} P(Y = 1|X) &= \frac{P(X|Y = 1)\pi}{P(X)} \\ &= \frac{1}{1 + \exp\big(- \beta_0 - \langle\beta_1, X\rangle\big)} \end{aligned}$$

Avec :

$$\begin{aligned} \beta_0 &= \log\big(\frac{\pi}{1 - \pi}\big) + \alpha_0/2 \\ \beta_1 &= \alpha_1/2 \end{aligned}$$

2.2

I love latex, but no, I'm not going to code it. This is just an adaptation of the proof in the example in dimention one of the Probabilistic Model class.

2.3

Everything worked fine and then I tried adding a λ_0... which broke everything and made the results unstable. The λ_0 is still there, but set to 0. You can try a non zero value to see what it does.

It can be seen as some sorte of regularization as it forces the covariance matrix to be somewhat close to a certain shape (diagonal).

2.4

A false positive/negative is when the model is wrongly predicting a positive/negative result, where it should be the opposite. The confusion matrix is not really interesting, there just seems to be a slight preference towards the -1 labels.

3 Question 3

I implemented the $K - means++$ algorithm. The PCA illustrates what we already said in this homework and the previous one : The A and C are clearly separable, but the B do not form a clear cluster and overlap with A and C. We can see it in the figure where points are colored by letter : the B are really spread out everywhere. The confusion matrix only reinforces this point, B is often confused with C and vis versa, as well as A, but A and C are never confused for the other.

A graphical explanation would be that B has both a large bar in the middle that is present in A, and a large bar/curve on the left that overlaps a lot with the curve of C.