4.20

Intro

In this exercise we will perform a brief theoretical comparision between **binary** generative and discriminative classifiers with different degrees of complexity (more specifically Gaussian Discriminant Analysis (GDA) and Logistic Regression (LogReg)). The exercise requires a careful reading of its statement since it is fairly long and contain a lot of details. In a few words, the first part of the exercise present us with four classifiers: 2 GDAs and 2 LogReg and a decision criteria L. The decision criteria is the **conditional log likelihood** on the **training set**. Furthermore, all the learners are trained using MLE. Here is important to mention that GDA has a close form solution, while LogReg do not have. Nevertheless, the latter gives rise to a convex optimization problem. We will assime that during training the **global optimum was found**.

In the second part of the exercise, the author present us with a new performance metrics R and ask us if conclusions made about the models with respect to L can be transfered to R. This is very important because is usually what happens in machine learning. In classification, we usually are interested in the accuracy score. However, our cost function is usually based on the MLE cross entropy, due to several reasons.

Before starting, I want to say that I am not 100% sure of my ansewer on item e, since I did not provide any rigorous mathematical argument. Nevertheless, I think my answer makes sense intuitively and holds for most cases. **If someone disagree** and have rigorous proof, feel free to contact me or post it on the GitHub page of the project :D.

Solution

a)

$$\begin{aligned} &GaussI = M_1 \\ &LinLog = M_2 \\ &L(M_1) \leq L(M_2) \end{aligned} \tag{1}$$

The explanation is the following: for LinLog, L is exactly what we are trying to optimize during training, because we are dealing with a discriminative classifier. Also, $p(y_i|x_i) = sigm(w^t x_i + w_0)$, where (w, w_0) is the weight vector which define the decision boundary plane on the input space x. Since this is a concave optimization problem (see Chapter 8 for proof), $(w, w_0)_{LogReg}$ is the global maximum that L can achieve when $p(y_i|x_i) = sigm(w^t x_i + w_0)$.

Meanwhile, for GaussI, the cost function during training is **not** equal to L. When dealing with generative classifiers, we maximize the joint log-likelihood:

$$\begin{aligned} &l(w, w_0) = \frac{1}{n} \sum_i log\ p(x_i, y_i|w, w_0) = \\ &\frac{1}{n} \sum_i log\ p(x_i|w, w_0) p(y_i|x_i, w, w_0) = \\ &\frac{1}{n} \sum_i log\ p(x_i|w, w_0) + L \end{aligned} \tag{2}$$

In section 4.2.3, the author showed that for LDA: $p(y_i|x_i) = sigm(w^t x_i + w_0)$. Therefore, the second term of the cost function of GaussI is equal to the cost

function of LogReg. So, we have two possibilities: either the solution w is equal for both classifiers, or the $\sum_i log\ p(x_i|w, w_0)$ term of GaussI cost function forces $w_{GaussI} \neq w_{LogReg}$. Either way, we have $L(M_1) \leq L(M_2)$

b)

$$\begin{aligned} &GaussX = M_1 \\ &QuadLog = M_2 \\ &L(M_1) \leq L(M_2) \end{aligned} \tag{3}$$

The reason is the same as in item a. The only difference is in this case, we have a quaratic function of the input x inside the sigmoid both for GaussX and QuadLog.

c)

$$\begin{aligned} &LinLog = M_1 \\ &QuadLog = M_2 \\ &L(M_1) \leq L(M_2) \end{aligned} \tag{4}$$

By setting the weights of the quadratic terms to 0, QuadLog becomes LinLog. Therefore, is not possible for LinLog to outperform QuadLog (remember that we are restraining ourself to the training set). On the other hand, it is possible that by using the weights of the quadratic terms, QuadLog outperforms LinLog.

d)

$$\begin{aligned} &GaussI = M_1 \\ &QuadLog = M_2 \\ &L(M_1) \leq L(M_2) \end{aligned} \tag{5}$$

This result is due to $L(M_{GaussI}) \leq L(M_{LinLog}) \leq L(M_{QuadLog})$.

e)

Since the name of the problem is 'Logistic Regression vs LDA/QDA', we will restric our models M to be one of the four mentioned. In this case, **I think** the answer is **true**. We can think about the answer geometrically: all the models are based on a decision boundary and a sigmoid function and the score L gets better as the instances are pushed to the correct side of the boundary (the right class) and distance themselves from the boundary limit (higher probabilities). Furthermore, due to the shape of the sigmoid function, its derivative is bigger around 0.5 than at the extremes where the function is saturated. Therefore, when we maximize L, which we do in all the models (in the GDA is just one of the terms of the cost function), it is more valuable to make a pass that decrease R (increase accuracy), than a pass that increases the distance between the boundary and the instances that already have the right class.

Conclusion

In this exercise, we made a theoretical comparision between the performance of different classifiers. We could feel that this task is not easy and we benefit a lot from the choice of performance L and the type of classifiers (GDA and LogReg). Trying to generalize beyond that is well studied in statistical learning theory and goes well beyond the scope of this book.

In item e, I do not know if my ansewer is the correct one, as I did not provide any rigorous mathematical argument. Nevertheless, the feeling behind it makes sense and as a general case, I think that this is correct. Once again, if anyone disagree and have a rigorous proof to do so, feel free to post on the GitHub page of the project.