Data Analysis PI
Theoretical assignment #7

Kupriyanov Kirill

Task 1.

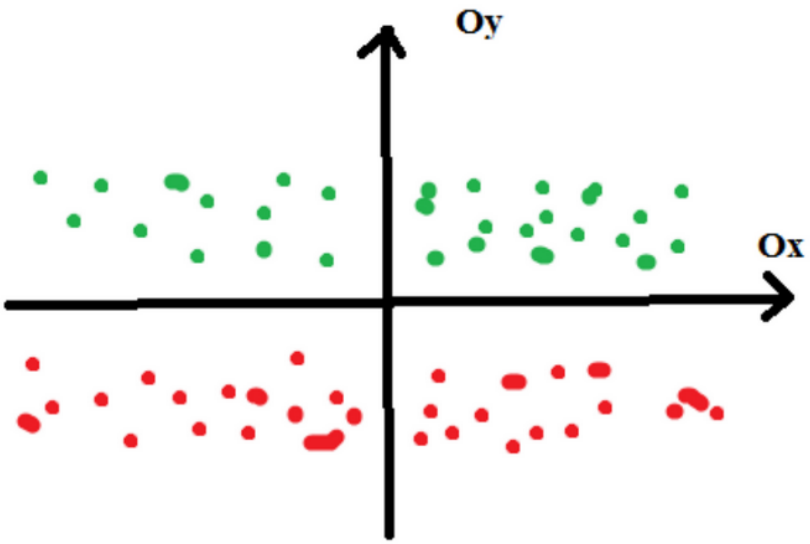
Problem: Does the principal component analysis (PCA) transformation require the preliminary feature standartization procedures (centering / scaling)? Explain your answer.

Solution: Перед использованием метода главных компонент предпочтительнее предварительно стандартизировать данные. Так как за основу метода взята дисперсия, масштабирование величин приведет к изменению главных компонент и как следствие результатов. Если какая-то величина будет измеряться гораздо большими числами, чем другая, то эта величина будет доминировать и, в основном, именно она будет влиять на результат. Это можно воспринимать как весы, и если такого в решаемой задаче не предполагается, то лучше данные стандартизировать.

Task 4.

Problem: Provide an example of the two-dimensional ($d = 2$) dataset in the binary classification problem for which the preliminary application of PCA compression to dimensionality $d = 1$ would hurt the classification accuracy dramatically. Explain, why PCA can hurt the classification accuracy?

Solution: Допустим у нас следующее распределение. Красным выделены объекты одного класса, зеленым другого.



В таком случае главная компонента будет примерно совпадать с осью Ox. Если мы спроектируем объекты на такую главную компоненту, то получим следующую картину:

В итоге, получилась смесь точек, которую невозможно адекватно классифицировать. Получается это в следствие того, что выборка вытянута по оси Ox, как следствие выбирается главная компонента схожая с Ox. Но ключевая информация о разделении классов находится на оси Oy. Иначе говоря, в этом примере, проектируя объекты, мы теряем всю информацию о классификации.