Тут титульник

# Содержание

# 1. Снижение размерности пространства признаков. Алгоритмы PCA и LDA

Метод главных компонент (principal component analysis, PCA) — один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.

Данный метод аппроксимирует n-размерное облако наблюдений до эллипсоида (тоже n-мерного), полуоси которого и будут являться будущими главными компонентами. И при проекции на такие оси (снижении размерности) сохраняется наибольшее количество информации.

Латентное размещение Дирихле (LDA, от англ. Latent Dirichlet allocation) — применяемая в машинном обучении и информационном поиске порождающая модель, позволяющая объяснять результаты наблюдений с помощью неявных групп, благодаря чему возможно выявление причин сходства некоторых частей данных. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем и что появление каждого слова связано с одной из тем документа.

В LDA каждый документ может рассматриваться как набор различных тематик. Подобный подход схож с вероятностным латентно-семантическим анализом (pLSA) с той разницей, что в LDA предполагается, что распределение тематик имеет в качестве априори распределения Дирихле. На практике в результате получается более корректный набор тематик.

# 2. Цель лабораторной работы

Цели:

Получить практические навыки по снижению входного пространства признаков

Задачи:

1. Применить к датасету Titanic алгоритм PCA.

2. Применить к датасету Titanic алгоритм LDA.

3. Провести ряд экспериментов используя данные из пунктов 1 и 2 и любой из классификаторов, рассмотренных в предыдущих работах.

# 3. Инструменты

В качестве инструментов для выполнения поставленной цели был выбран язык Python и библиотеки scikit-learn и Pandas. Бибилотека Pandas была использована для подготовки датасета к будущему использованию.

Для реализации алгоритмов PCA и LDA использовался класс PCA из sklearn.decomposition и класс LinearDiscriminantAnalysis из sklearn.discriminant_analysis.

Основные параметры класса PCA.

n_components – количество главных компонент.

# 4. Эксперименты

Для того, чтобы вывести при каких параметрах модель LinearDiscriminantAnalysis и PCA дает наиболее точный прогноз, необходимо провести ряд экспериментов с разными значениями параметров.

Для того, чтобы оценить качество уменьшения размерности PCA и качество категоризации LDA, обучим на полученных данных модель RandomForestClassifier из предыдущей лабораторной работы. Для построения "случайного леса"используем те параметры, при которых в прошлой лабораторной мы получили наилучшие результаты.

Таблица 1 - Точность прогнозов при использовании PCA.

| № | Параметры | Точность (в процентах) |
|---|---|---|
| 1 | n_components=2 | 60.422 |

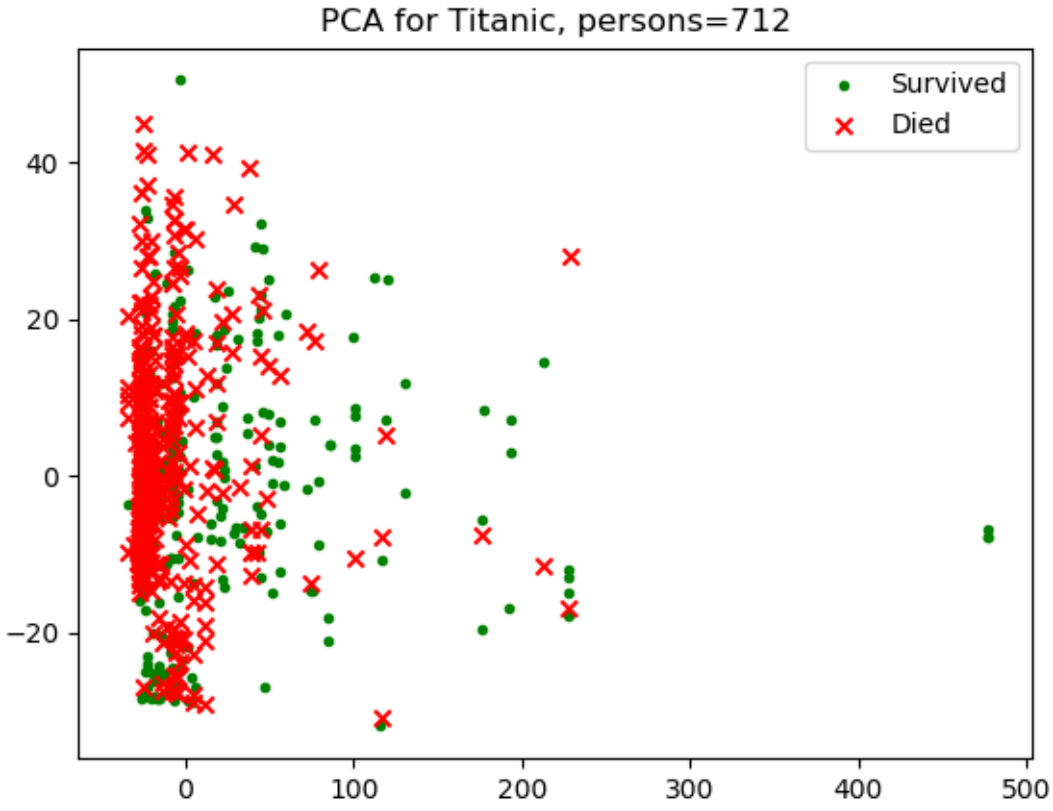
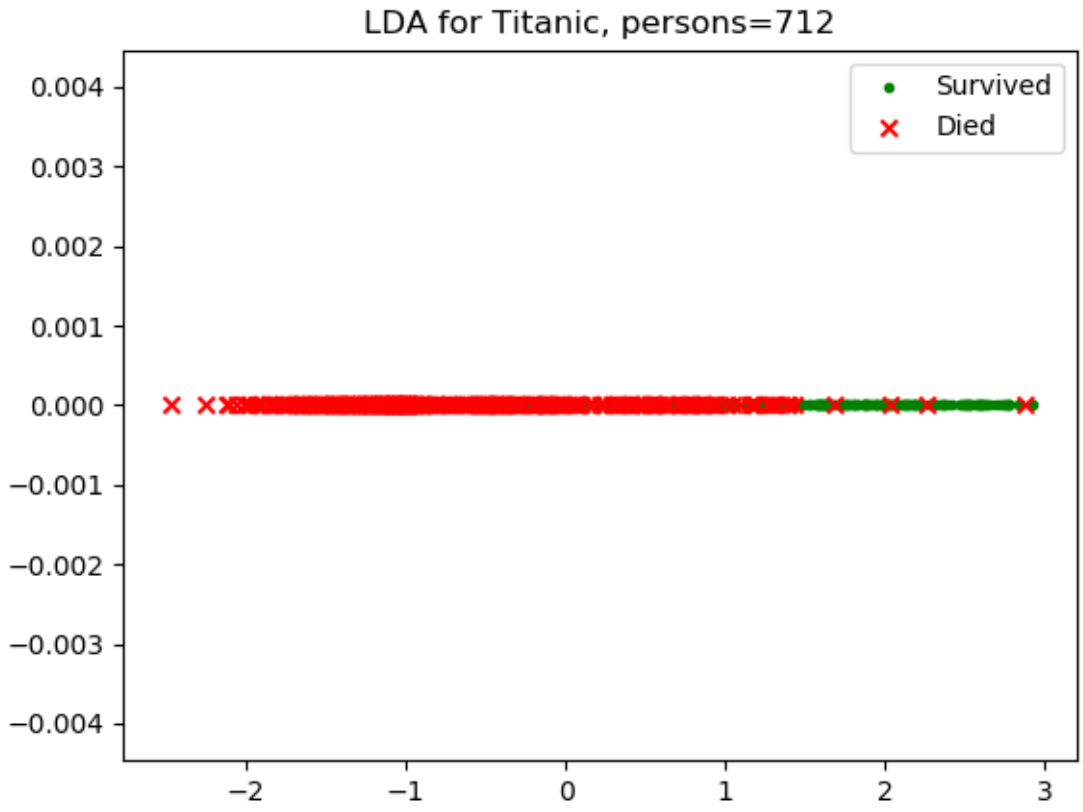



Таблица 2 - Точность прогнозов при использовании LDA.

| № | Параметры | Точность (в процентах) |
|---|---|---|
| 1 | n_components=2 | 64.652 |



# 5. Итог

В ходе проведения экпериментов были выявлено, что при изменении данных с помощью алгоритмов PCA и LDA произошло ухудшение результатов работы алгоритма случайного леса. Это может быть связано со структурой набора данных. Так как данные, вероятно, коррелируют между собой, можно предположить, что это сказывается на результатах изменения размерности.

Также, эксперименты с LDA показали, что параметры модели не оказывают значительного влияния. Поэтому в таблице запечатлен прогноз с параметрами по умолчанию.

В случае PCA с параметрами по умолчанию удалось получить наихудший результат. Лучший был получен при параметре svd_solver равном randomized и большом числе компонент. Исходя из других опытов, можно сделать вывод, что svd_solver на самом деле не играет большой роли, самым важным параметром является n_components - размерность выходных данных.