Отбор признаков и линейные методы снижения размерности

Кантонистова Елена

elena.kantonistova@yandex.ru

27 апреля 2019

План лекции

1. Методы отбора признаков
   - VarianceThreshold
   - Отбор по корреляции с целевой переменной
   - Более сложные методы отбора признаков

2. Линейные методы снижения размерности
   - Метод главных компонент
   - Линейный дискриминантный анализ

VarianceThreshold

Можем удалить признаки, которые имеют очень маленькую дисперсию, т.е. практически константы.

Отбор по корреляции с целевой переменной

Для каждого признака вычислим его корреляцию с целевой переменной. Будем выкидывать признаки, имеющие маленькую корреляцию.

Более сложные методы отбора признаков

- Filtration methods (фильтрационные методы)
- Wrapping methods (оберточные методы)
- Model selection (встроенный в модель отбор признаков)

Фильтрационные методы

Фильтрационные методы - это отбор признаков по различным статистическим тестам. Идея метода состоит в вычислении влияния каждого признака в отдельности на целевую переменную (с помощью вычисления некоторой статистики).
Очевидный плюс метода: скорость, так как мы вычисляем значения N статистик, где N - количество признаков.

Фильтрационные методы

В sklearn есть сразу несколько методов, использующих отбор по статистическим критериям. Среди них выделим следующие:

- SelectKBest - оставляет k признаков с наибольшим значением выбранной статистики
- SelectPercentile - оставляет признаки со значениями выбранной статистики, попавшими в заданную пользователем квантиль
- и другие (см.sklearn)

Статистические тесты для отбора признаков

- mutual information: для векторов X и Y статистика вычисляется по формуле

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

- хи-квадрат:

$$\chi^2(X,Y) = \sum_{i=1}^{n} \frac{(Y_i - X_i)^2}{X_i}$$

- f-regression - тест, основанный на корреляции линейного регрессора с целевой переменной

Оберточные методы

Оберточные методы используют жадный отбор признаков, т.е. последовательно выкидывают наименее подходящие по мнению методов признаки.
В sklearn есть оберточный метод - Recursive Feature Elimination (RFE).
Параметры метода:
a) алгоритм, используемый для отбора признаков (например, RandomForest)
b) число признаков, которое мы хотим оставить.

Встроенные в модель методы

При добавлении регуляризатора в модель мы уменьшаем влияние различных признаков на финальную модель. В случае, если мы добавляем l_1-регуляризатор, то в силу вида регуляризатора (сумма модулей весов), модель автоматически обнуляет веса некоторых признаков, то есть выкидывает их.
Линейная модель с l_1-регуляризатором в sklearn: LASSO.

Снижение размерности

Предыдущие методы отбирали из исходных признаков некоторое подмножество признаков. Теперь мы хотим придумать новые признаки, каким-то образом выражающиеся через старые, причем новых признаков хочется меньше, чем старых. Сегодня будем рассматривать только случай, когда новые признаки линейно выражаются через старые.

Метод главных компонент

Постановка задачи:
$f_1(x), ..., f_n(x)$— исходные числовые признаки;
$g_1(x), ..., g_m(x)$ — новые числовые признаки, $m \leq n$;

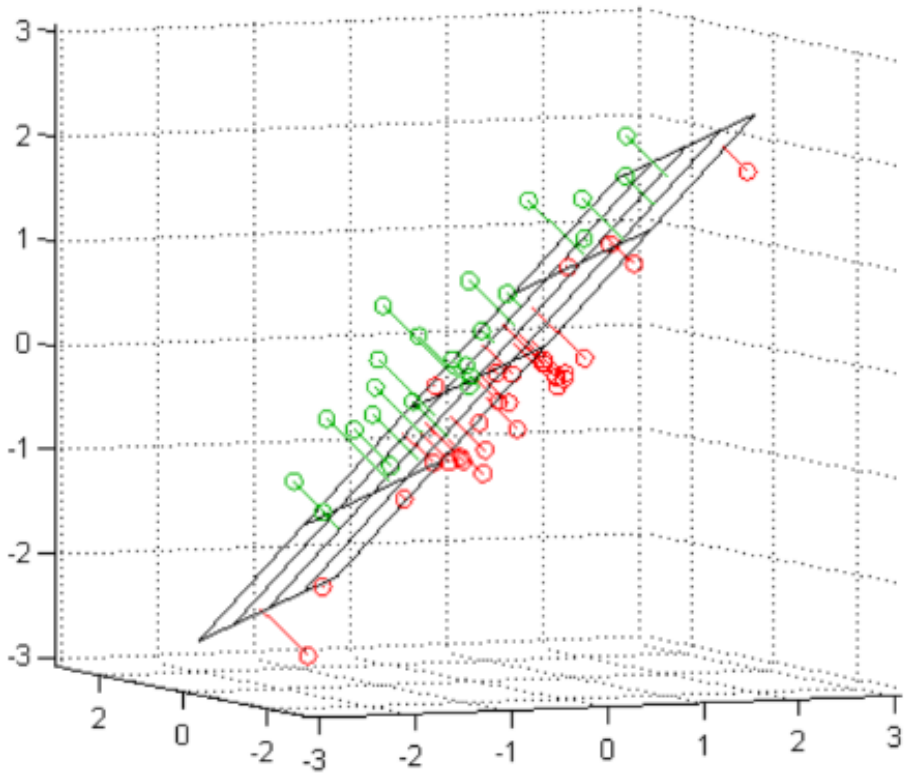
Мы хотим, чтобы новые числовые признаки $g_i(x)$ линейно выражались через исходные признаки $f_j(x)$, при этом чтобы исходные признаки также линейно восстанавливались по новым признакам. При этом мы хотим, чтобы при переходе к новым признакам было потеряно наименьшее количество исходной информации.

Метод главных компонент

Метод главных компонент работает только с признаками. Для него не важна целевая переменная (если она есть). Таким образом, метод главных компонент - это обучение без учителя.

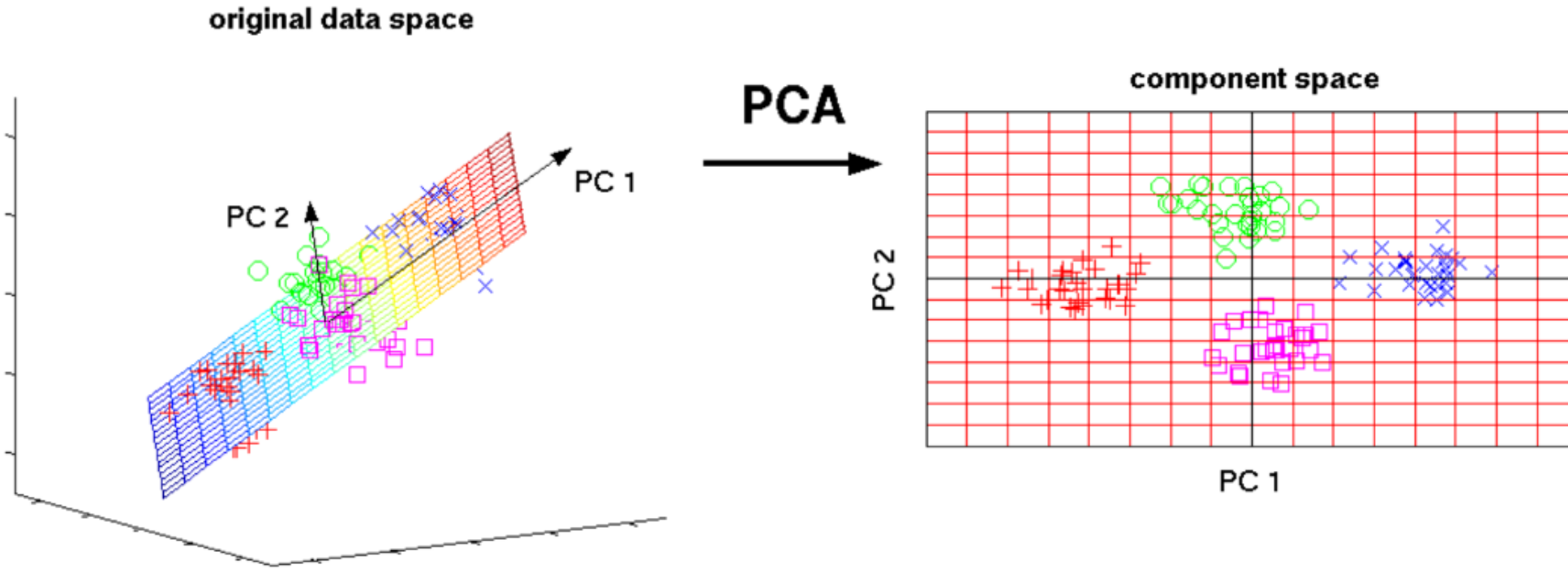
Геометрическая интерпретация PCA

Геометрически метод главных компонент ищет гиперплоскость заданной размерности, при проекции на которую сумма квадратов расстояний от исходных точек будет минимальной.



Визуализация проекции на гиперплоскость

Точки, плохо разделимые в исходном пространстве, могут быть лучше разделимы при проекции на некоторую гиперплоскость.



Метод главных компонент

Перед применением метода необходимо центрировать данные, то есть вычесть из каждого признака его среднее значение.

- Пусть X - матрица объект-признак
- Метод главных компонент осуществляет проекцию исходных объектов на гиперплоскость некоторой размерности d.

Теорема. Базисные векторы этой гиперплоскости - это собственные векторы матрицы $X^T X$, соответствующие d её наибольшим собственным значениям.

Доля объясненной дисперсии

- Упорядочим собственные значения матрицы $X^T X$ по убыванию: $\lambda_1 \geq \lambda_2 \geq \cdots > \lambda_n \geq 0$.

- Доля дисперсии, объяснённой j-й компонентой (explained variance ratio):

$$\delta_j = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_n}$$

- Доля дисперсии, объясняемой первыми k компонентами:

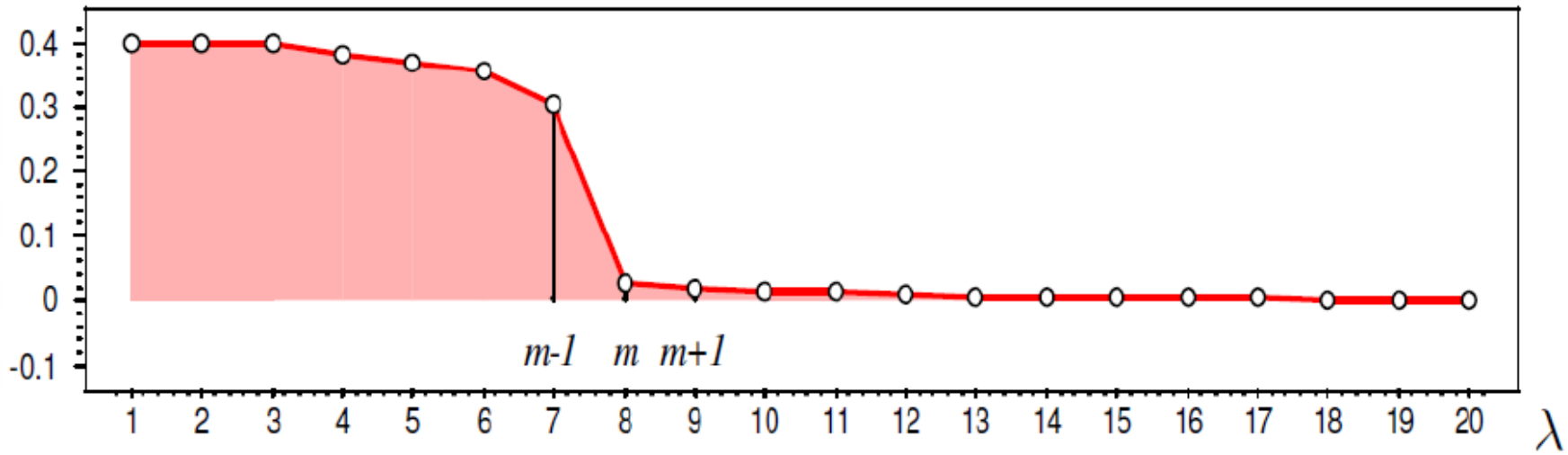
$$\delta = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_n}$$

Выбор числа главных компонент

- Эффективная размерность выборки – это наименьшее целое m, при котором

$$E_m = \frac{||ZU^T - X||^2}{||X||^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\sum_{i=1}^{n} \lambda_i} \leq \varepsilon$$

Критерий крутого склона:



Faces dataset

Faces dataset (main components)

Первые главные компоненты после применения PCA

Восстановленное изображение

Linear Discriminant Analisys, LDA

LDA - это обучение с учителем. При помощи метода линейного дискриминантного анализа выбирается проекция исходного пространства признаков на новое пространство признаков таким образом, чтобы минимизировать внутриклассовый разброс точек и максимизировать межклассовое расстояние в пространстве признаков.

LDA, формализация

- Классификация между ω_1 и ω_2.
- Пусть $C_1 = \{i : x_i \in \omega_1\}, \quad C_2 = \{i : x_i \in \omega_2\}$ и

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n, \quad m_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

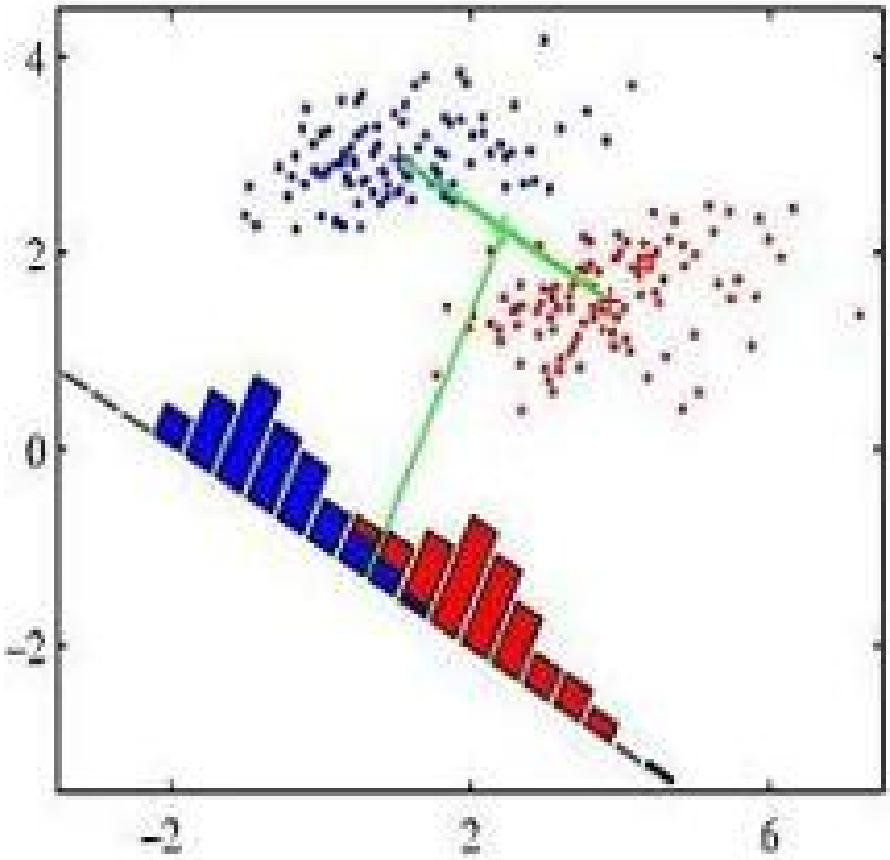
$$\mu_1 = w^T m_1, \quad \mu_2 = w^T m_2$$

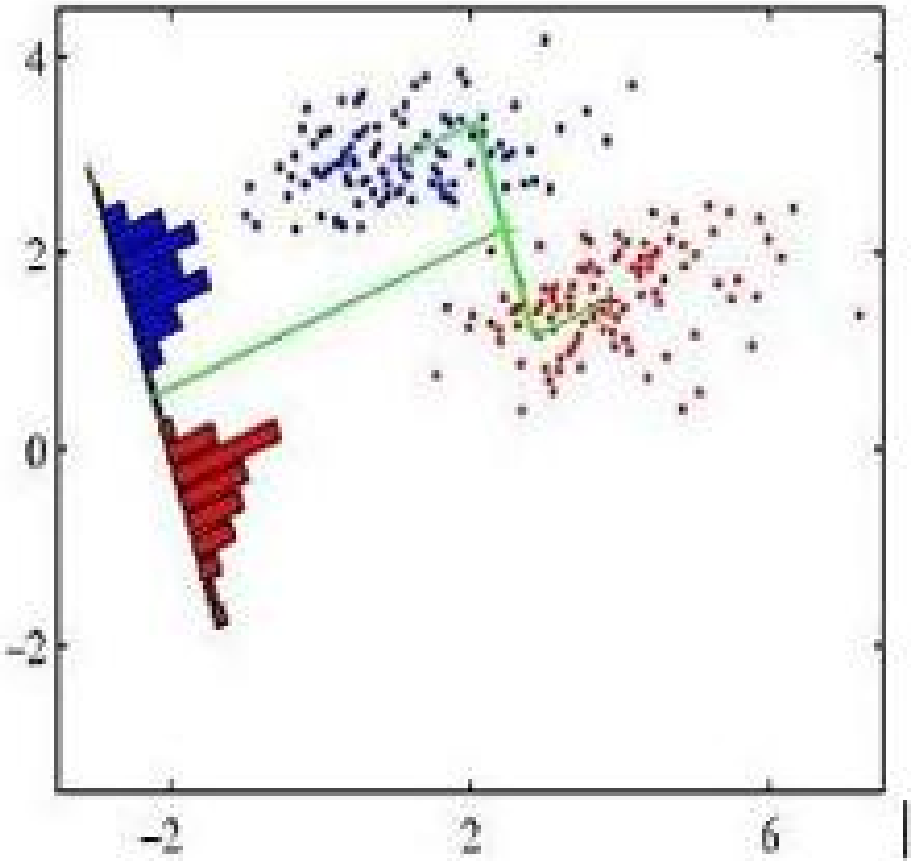
- Определим дисперсии спроецированных на подпространство w классов:

$$s_1 = \sum_{n \in C_1} (w^T x_n - w^T m_1)^2, \quad s_2 = \sum_{n \in C_2} (w^T x_n - w^T m_2)^2$$

- Критерий LDA Фишера: $\frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2} \to \max_w$

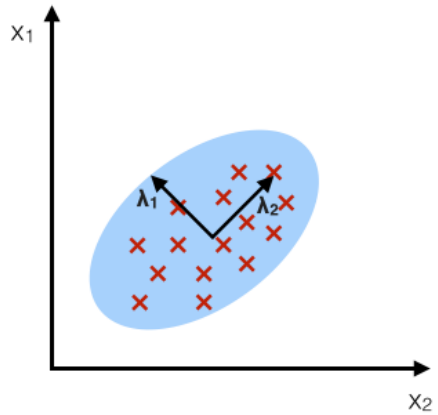
LDA, пример





PCA vs LDA

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation