МИНОБРНАУКИ РОССИИ

РГУ нефти и газа (НИУ) имени И.М. Губкина

|  |  |
| --- | --- |
| Факультет | **Автоматики и Вычислительной техники** |
| Кафедра | **Информатики** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Оценка комиссии: |  | | | Рейтинг: |  |
| Подписи членов комиссии: | | | | | |
|  | |  | Вишневская Е. А. | | |
| (подпись) | |  | (фамилия, имя, отчество) | | |
|  | |  |  | | |
| (подпись) | |  | (фамилия, имя, отчество) | | |
|  | | | | | |
| (дата) | | | | | |
|  | |  |  | | |

**ДОКЛАД**

|  |  |
| --- | --- |
| по дисциплине | Средства интеллектуального анализа данных и машинное |
| обучение | |

|  |  |
| --- | --- |
| на тему | PCA – метод главных компонент |
|  | |
|  | |

|  |  |  |  |
| --- | --- | --- | --- |
| «К ЗАЩИТЕ» |  | ВЫПОЛНИЛ: |  |
|  |  | Студент группы | **АА-22-08** |
|  |  |  | (номер группы) |
| Доцент, к. ф. м. н.: Вишневская Е. А. |  | Сафуанов Артур Ришатович  Коргин Артём Денисович | |
| (должность, ученая степень; фамилия, и.о.) |  | (фамилия, имя, отчество) | |
|  |  |  | |
| (подпись) |  | (подпись) | |
|  |  |  | |
| (дата) |  | (дата) | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Москва, 20 | 25 |  |

**1. Введение в PCA: Суть и Цели**

Метод Главных Компонент (Principal Component Analysis, PCA) – это фундаментальный статистический метод, относящийся к области анализа данных и машинного обучения. Его основное предназначение – **сокращение размерности** данных. Это означает уменьшение количества признаков (переменных), описывающих каждое наблюдение в наборе данных, при одновременном сохранении максимально возможного объема полезной информации, содержащейся в исходных данных.

Ключевая идея PCA заключается в преобразовании исходного пространства признаков. Вместо работы с исходными, потенциально коррелирующими и избыточными признаками, PCA находит новый набор ортогональных (независимых) признаков, называемых **Главными Компонентами (Principal Components, PC)**. Эти компоненты представляют собой линейные комбинации исходных признаков и располагаются в порядке убывания их значимости: первая главная компонента (PC1) объясняет наибольшую долю изменчивости (дисперсии) данных, вторая (PC2) – наибольшую долю оставшейся изменчивости (при условии ортогональности к первой), и так далее.

**2. Задачи, решаемые с помощью PCA**

Сокращение размерности с помощью PCA решает несколько важных практических задач:

1. **Улучшение Визуализации Данных:** наиболее очевидное применение. Данные, имеющие десятки или сотни признаков, невозможно наглядно отобразить. PCA позволяет спроецировать данные на 2 или 3 первые главные компоненты, объясняющие основную часть дисперсии. Это дает возможность построить информативные двумерные или трехмерные графики (диаграммы рассеяния) для визуального исследования структуры данных, выявления кластеров, аномалий или закономерностей.
2. **Сокращение Вычислительной Сложности (Ускорение Работы Алгоритмов):** многие алгоритмы машинного обучения (особенно основанные на расстояниях, такие как k-NN, или требующие обращения больших матриц) страдают от "проклятия размерности" – их производительность резко падает с ростом числа признаков. Уменьшение количества признаков с помощью PCA (выбор k << d, где d – исходная размерность) приводит к значительному сокращению времени обучения моделей и потребления вычислительных ресурсов (памяти, процессорного времени).
3. **Улучшение Качества Моделей Машинного Обучения:** связано с предыдущим пунктом. Уменьшение размерности не только ускоряет обучение, но и может повысить обобщающую способность моделей. Избыточные или шумовые признаки могут ухудшать качество модели. PCA, сохраняя основную информацию и отбрасывая шум/избыточность, помогает построить более устойчивые и точные модели (например, классификаторы или регрессоры). Это особенно актуально при наличии мультиколлинеарности (сильной корреляции между исходными признаками).
4. **Выявление Скрытых Закономерностей и Зависимостей:** анализируя главные компоненты и их "веса" (коэффициенты линейной комбинации исходных признаков), можно понять, какие исходные признаки вносят наибольший вклад в наблюдаемую изменчивость данных и как они взаимосвязаны. Это помогает обнаружить скрытые факторы или структуры, неочевидные при анализе исходных признаков по отдельности. Компоненты могут иметь интерпретируемый смысл (например, "общий размер", "форма").

**3. Математические Основы и Алгоритм PCA**

В основе PCA лежит линейная алгебра и анализ ковариационной матрицы данных. Алгоритм можно представить в виде следующих шагов:

**Шаг 1: Стандартизация Данных**

* **Цель:** привести все признаки к единому масштабу (среднее = 0, стандартное отклонение = 1). Это критически важно, так как PCA чувствителен к масштабу переменных. Признак с большим разбросом значений будет искусственно вносить больший вклад в дисперсию и доминировать при построении главных компонент, если не выполнить стандартизацию.
* **Действия:** загрузить данные. Разделить выборку на матрицу признаков X (где каждая строка - наблюдение, каждый столбец - признак) и вектор целевой переменной (target), если она есть (для PCA сама target обычно не используется на этом этапе, но может быть нужна для последующего анализа). Выполнить предподготовку данных (обработка пропусков, выбросов – если требуется). Рассчитать среднее значение и стандартное отклонение для каждого признака. Преобразовать каждый признак: X\_std = (X - mean(X)) / std(X).

**Шаг 2: Построение Ковариационной Матрицы**

* **Цель:** вычислить матрицу, описывающую попарные ковариации между всеми стандартизированными признаками. Ковариация показывает направление линейной зависимости между двумя признаками.
* **Действия:** для стандартизированной матрицы данных X\_std размером n x d (n - число наблюдений, d - число признаков) ковариационная матрица C вычисляется как: C = (1/(n-1)) \* X\_std^T \* X\_std. Это симметричная матрица размером d x d. Элемент C[i,j] равен ковариации между i-м и j-м признаками.

**Шаг 3: Сингулярное Разложение (SVD) или Вычисление Собственных Векторов и Собственных Значений**

* **Цель:** найти направления (собственные векторы) в исходном пространстве признаков, вдоль которых дисперсия данных максимальна, и величину этой дисперсии (собственные значения).
* **Действия:** для ковариационной матрицы C решается задача нахождения собственных значений (λ) и собственных векторов (v): C \* v = λ \* v. Собственные векторы v и есть направления главных компонент. Соответствующие им собственные значения λ пропорциональны доле дисперсии данных, объясняемой каждой главной компонентой (чем больше λ, тем больше дисперсии объясняет компонента).

**Шаг 4: Сортировка Собственных Значений и Векторов**

* **Цель:** упорядочить главные компоненты по их значимости (объясняемой дисперсии).
* **Действия:** собственные значения сортируются в порядке убывания: λ1 >= λ2 >= ... >= λd. Соответствующие им собственные векторы v1, v2, ..., vd переупорядочиваются в том же порядке. Вектор v1 задает направление первой (самой важной) главной компоненты PC1, v2 – направление PC2, и т.д.

**Шаг 5: Выбор Количества Главных Компонент k**

* **Цель:** определить, сколько главных компонент k (где k < d, часто k << d) достаточно для представления данных без существенной потери информации.
* **Действия:** есть несколько подходов:
  + **Объясненная дисперсия:** рассчитывается доля объясненной дисперсии для каждой компоненты: λi / Σ(λj) (для j=1..d). Выбирают k так, чтобы суммарная доля объясненной дисперсии первых k компонент превышала заданный порог (например, 80%, 90% или 95%). В примере из презентации 2 компоненты объясняли около 60% дисперсии.
  + **График Scree Plot:** строится график собственных значений в порядке убывания. Точка, где график образует "локоть" (резкий перегиб), часто указывает на разумное значение k.
  + **Компромисс:** выбор k всегда является компромиссом между степенью сжатия данных (вычислительная эффективность), сохранением информации (производительностью последующих моделей) и необходимостью визуализации (часто k=2 или k=3).

**Шаг 6: Построение Матрицы Проекции**

* **Цель:** сформировать матрицу для преобразования данных в новое пространство меньшей размерности.
* **Действия:** из отсортированных собственных векторов выбираются первые k векторов (v1, v2, ..., vk). Эти векторы объединяются по столбцам в матрицу проекции W размером d x k. Столбцы матрицы W – это направления новых осей (главных компонент).

**Шаг 7: Преобразование Данных (Проекция в новое пространство)**

* **Цель:** получить представление исходных данных в пространстве k главных компонент.
* **Действия:** стандартизированная матрица данных X\_std умножается на матрицу проекции W: Y = X\_std \* W. Результирующая матрица Y размером n x k и содержит новые признаки – координаты каждого исходного наблюдения в системе k главных компонент. Это и есть данные в сокращенной размерности.

**4. Визуализация и Интерпретация Результатов PCA**

После выполнения PCA, особенно при выборе k=2 или k=3, данные можно визуализировать на диаграмме рассеяния. Как показано в презентации:

1. **Оси:** ось x соответствует первой главной компоненте (PC1), ось Y – второй главной компоненте (PC2).
2. **Разброс Данных:** наблюдения наносятся на график в соответствии с их координатами по PC1 и PC2.
3. **Интерпретация Дисперсии:** данные обычно демонстрируют больший разброс вдоль оси PC1, чем вдоль PC2. Это прямое следствие того, что PC1 объясняет наибольшую долю дисперсии данных. Чем больше разброс вдоль оси, тем больше информации (вариабельности) несет соответствующая компонента. В примере презентации PC1 объясняла больше данных, чем PC2, поэтому точки сильнее "растянуты" вдоль оси X.
4. **Выявление Структур:** такая визуализация позволяет увидеть кластеры наблюдений, выбросы, возможные линейные или нелинейные зависимости между главными компонентами.

**5. Заключение**

Метод Главных Компонент (PCA) является мощным и широко применяемым инструментом в анализе данных и машинном обучении. Его основная сила заключается в способности эффективно снижать размерность сложных многомерных наборов данных, сохраняя при этом основную информацию. Это достигается за счет построения нового набора некоррелированных признаков (главных компонент), упорядоченных по степени объясняемой ими дисперсии исходных данных.

Алгоритм PCA, основанный на линейно-алгебраических операциях со ковариационной матрицей (стандартизация, вычисление собственных векторов/значений, проекция), предоставляет систематический подход к выполнению этого преобразования. Практические преимущества PCA включают улучшенную визуализацию данных, значительное ускорение работы алгоритмов за счет сокращения признаков, потенциальное повышение качества моделей машинного обучения за счет борьбы с проклятием размерности и удаления шума, а также возможность выявления скрытых структур и зависимостей в данных.

Выбор оптимального числа главных компонент k является важным практическим решением, основанным на анализе объясненной дисперсии и требуемом балансе между сжатием данных и сохранением информации. Визуализация проекции данных на первые две или три главные компоненты служит не только для наглядности, но и для первичного исследования структуры данных.

Таким образом, PCA остается фундаментальным методом предобработки данных и извлечения признаков, необходимым в арсенале любого специалиста по данным.