Лабораторная работа №8
Классификация (линейный дискриминантный анализ, метод опорных векторов)

Цель:

Ознакомиться с методами кластеризации модуля Sklearn

- Контрольные вопросы

Выполнение:
1. Загрузка данных:

1.1. Загрузить датасет по ссылке: https://archive.ics.uci.edu/dataset/109/wine.

- Данные представлены в виде data файла. Данные представляют собой информацию о **трех** классах вина.
- Необходимо прочитать описание к данным (Variables Table) и определить, какие столбцы относятся к данным, а какие к меткам данных. Метки данных не используются для анализа данных и исключаются из набора

> **Дать пояснение**
> 1. Перечислить столбец (или столбцы, если их несколько), которые отнесены к меткам

1.2. Создать Python скрипт. Загрузить данные в датафрейм

1.3. Выделить в отдельные переменные данные (например, в `X`) и их метки (например, в `labels`)

1.4. Преобразовать метки в `число` с использованием LabelEncoder и LabelEncoder.fit_transform

- См. пример Label encoding

1.5. Разделить данные на на обучающую (`x_train`,`y_train` - 50% данных) и тестовую выборки (`x_test`,`y_test` - 50% данных)

- Используется train_test_split
- См. примеры кода в 3.1. Cross-validation: evaluating estimator performance

2. Линейный дискриминантный анализ

2.1. Провести классификацию наблюдений используя метод LinearDiscriminantAnalysis

- Пример Linear and Quadratic Discriminant Analysis, Линейный Дискриминантный Анализ и Квадратичный Дискриминантный Анализ

- Wiki Линейный дискриминантный анализ
- Линейный дискриминантный анализ (LDA). Принцип работы и реализация с нуля на Python / Хабр
- Линейный дискриминантный анализ (с примерами)
- Вывести на экран общее количество точек и количество неправильно маркированных точек
- Вычислить точность классификации (функция LinearDiscriminantAnalysis.score))

> **Дать пояснение**
> 1. Расшифровать атрибуты функции, на что они влияют?

2.2. Постройте два графика - зависимость неправильно классифицированных наблюдений и зависимость точности классификации от размера тестовой выборки.

- Размер тестовой выборки (`test_size`) изменяйте от 0.05 до 0.95 с шагом 0.05.
- Параметр `random_state` сделайте равным номеру своей зачетной книжки.

> **Дать пояснения**
> 1. Что происходит с результатами с изменением размера тестовой выборки?

2.3. Выполните преобразование данных с помощью функции LinearDiscriminantAnalysis.transform

- Визуализируйте результат работы функции `LinearDiscriminantAnalysis.transform`
  * Пример использования Comparison of LDA and PCA 2D projection of Iris dataset

> **Дать пояснения**
> 1. Какие действия выполняет функция `LinearDiscriminantAnalysis.transform`?

2.4. Постройте по два графика - зависимость неправильно классифицированных наблюдений и зависимость точности классификации от размера тестовой выборки для различных параметров solver(svd, lsqr, eigen), shrinkage(auto, None)

- Размер тестовой выборки (`test_size`) изменяйте от 0.05 до 0.95 с шагом 0.05.
- Параметр `random_state` сделайте равным номеру своей зачетной книжки.

> **Дать пояснения**
> 1. Чем отличаются полученные результаты?
> 2. Как влияют различные параметры на получаемый результат?

2.5. Задайте априорную вероятность классу с номером 1 равную 0.6, остальным классам задайте равные априорные вероятности.

- Значение вероятности необходимо задавать в параметре `priors` для каждого класса*

> **Дать пояснения**
> 1. Какова суммарная вероятность для всех классов?
> 2. Как изменился результат?

3. Метод опорных векторов

3.1. Проведите классификацию этих же данных одним из методов опорных векторов (SVC)

- Все методы опорных векторов из библиотеки sklearn - sklearn - vector machine algorithms
- Метод опорных векторов (Support Vector Machines - SVM)
- Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) / Хабр
- Пример Support Vector Machines, Модели для классификации: Метод опорных векторов (SVM), Метод опорных векторов (SVC, SVR)
- Вывести на экран общее количество точек и количество неправильно маркированных точек
- Вычислить точность классификации (функция SVC.score))
- Выведите на экран значение `SVC.support_vectors_`, `SVC.support_`, `SVC.n_support_`

> **Дать пояснения**
> 1. Что означают выведенные значения?
> 2. Как исходные данные и параметры классификации влияют на эти значения?

3.2. Постройте два графика - зависимость неправильно классифицированных наблюдений и зависимость точности классификации от размера тестовой выборки.

- Размер тестовой выборки (`test_size`) изменяйте от 0.05 до 0.95 с шагом 0.05.
- Параметр `random_state` сделайте равным номеру своей зачетной книжки.

> **Дать пояснения**
> 1. Что происходит с результатами с изменением размера тестовой выборки?

3.3. Постройте по два графика - зависимость неправильно классифицированных наблюдений и зависимость точности классификации от размера тестовой выборки для различных параметров kernel('linear', 'poly', 'rbf', 'sigmoid', 'precomputed'), degree(от 3 до 10), max_iter(-1,40,200,400)

- Размер тестовой выборки (`test_size`) изменяйте от 0.05 до 0.95 с шагом 0.05.
- Параметр `random_state` сделайте равным номеру своей зачетной книжки.

> **Дать пояснения**
> 1. Чем отличаются полученные результаты?
> 2. Как влияют различные параметры на получаемый результат?

3.4. Повторите пункт **3.1.** с использованием методов NuSVC и LinearSVC

> **Дать пояснения**
> 1. Чем отличаются методы NuSVC и LinearSVC от метода SVC?
> 2. Чем отличаются полученные результаты от метода SVC?

Контрольные вопросы

Контрольные вопросы по LinearDiscriminantAnalysis (LDA)

Эти вопросы помогут вам проверить свои знания о LDA и его применении в задачах классификации и понижения размерности.

I. Основы и принципы работы:

1. Что такое LDA? Опишите основные принципы работы LDA для классификации.
2. Чем LDA отличается от PCA (Principal Component Analysis)? В чем их сходства и различия?
3. Как LDA использует информацию о классах при понижении размерности?
4. Какие предположения делает LDA о данных? Как нарушение этих предположений может повлиять на производительность?
5. Как LDA рассчитывает дискриминантные функции?
6. Что такое within-class scatter matrix и between-class scatter matrix? Как они используются в LDA?
7. Как LDA определяет оптимальное направление проекции для разделения классов?

II. Гиперпараметры и их влияние:

1. Какие основные гиперпараметры есть у LDA в scikit-learn? (`solver`, `shrinkage`, `n_components`)
2. Как выбор `solver` влияет на вычисления в LDA? Когда следует использовать каждый из доступных solvers?
3. Что такое `shrinkage` и как он помогает улучшить производительность LDA, особенно при большом количестве признаков или малом объеме данных?
4. Как `n_components` влияет на размерность преобразованных данных?

III. Применение и практика:

1. Как использовать LinearDiscriminantAnalysis в scikit-learn? Приведите пример кода для обучения и предсказания.
2. Как оценить качество модели LDA? Какие метрики можно использовать?
3. Как LDA обрабатывает многоклассовую классификацию?
4. Как можно использовать LDA для визуализации данных?
5. Как обрабатывать пропущенные значения при использовании LDA?

IV. Преимущества и недостатки:

1. Каковы преимущества использования LDA для классификации и понижения размерности?
2. Каковы недостатки LDA? В каких случаях он может быть неэффективен?

V. Более сложные вопросы:

1. Как LDA связан с линейным регрессионным анализом?
2. Как можно использовать LDA для выбора признаков?
3. Как влияет корреляция между признаками на производительность LDA?

4. Как можно обобщить LDA на нелинейные случаи (например, Kernel LDA)?
5. Когда предпочтительнее использовать LDA вместо других методов понижения размерности, таких как PCA или t-SNE?

Контрольные вопросы по SVC, NuSVC и LinearSVC:

Общие вопросы для SVC, NuSVC и LinearSVC:

1. Что такое метод опорных векторов и какова его основная идея?
2. Что такое разделяющая гиперплоскость и отступ (margin)? Почему важно максимизировать отступ?
3. Что такое опорные векторы и почему они важны?
4. Как SVM работает с линейно неразделимыми данными? Объясните концепцию ядерного трюка и ядерных функций.
5. Как параметр регуляризации (C в SVC и LinearSVC, nu в NuSVC) влияет на модель? Что происходит при слишком большом или слишком малом значении этого параметра?
6. Как оценить качество модели SVM? Какие метрики можно использовать?
7. Как SVM обрабатывает многоклассовую классификацию?
8. Как обрабатывать пропущенные значения при использовании SVM?

SVC:

1. Какие типы ядерных функций поддерживает SVC? Приведите примеры и опишите их применение (линейное, полиномиальное, RBF, сигмоидальное, кастомное).
2. Как параметры `gamma`, `degree` и `coef0` влияют на различные ядерные функции?
3. Как выбрать подходящее ядро для конкретной задачи?

NuSVC:

1. Чем NuSVC отличается от SVC? Что представляет собой параметр `nu` и как он связан с параметром `C` в SVC?
2. Какие преимущества и недостатки использования NuSVC по сравнению с SVC?
3. В каких случаях NuSVC может быть предпочтительнее SVC?

LinearSVC:

1. Чем LinearSVC отличается от SVC с линейным ядром?
2. Какие оптимизации использует LinearSVC, что делает его более эффективным для больших наборов данных?
3. Какие гиперпараметры специфичны для LinearSVC (`loss`, `penalty`, `dual`)? Как они влияют на модель?
4. В чем разница между `hinge` и `squared hinge` loss?
5. Когда следует использовать LinearSVC вместо SVC с линейным ядром?

Сравнение:

1. Сравните SVC, NuSVC и LinearSVC. В каких случаях следует использовать каждый из них?
2. Какие преимущества и недостатки есть у каждого из этих классификаторов?
3. Как выбрать наиболее подходящий классификатор SVM для конкретной задачи?