

Липецкий государственный технический университет

Институт компьютерных наук
Кафедра прикладной математики и системного анализа

Отчет по лабораторной работе № 3
по дисциплине «Интеллектуальные методы анализа данных»
на тему «Сравнение методов классификации»

Студент

Группа ПМ-23-1

Руководитель

к.т.н., доцент

учёная степень, учёное звание

подпись, дата

Габбасов.Д.В

фамилия, инициалы

Сысоев А.С.

фамилия, инициалы

Липецк 2025 г.

Постановка задачи

Цель работы

Освоить методы построения и сравнения различных типов классификаторов: линейных, нелинейных и байесовских подходов. Изучить особенности применения линейного и квадратичного дискриминантного анализа.

Задачи

1. Изучить основы методов классификации и их применение для решения задач распознавания образов.
2. Освоить построение линейных классификаторов.
3. Изучить методы нелинейной классификации.
4. Освоить байесовские методы классификации.
5. Изучить применение линейного и квадратичного дискриминантного анализа.
6. Сравнить качество различных моделей классификации.

Исходные данные

Выбрать один из датасетов:

- Вариант 1: Breast Cancer Wisconsin Dataset
- Вариант 2: Iris Plants Dataset
- Вариант 3: Wine Recognition Dataset
- Вариант 4: Pima Indians Diabetes Dataset

Оглавление

1. Введение	4
2. Основная часть	8
2.1. Подготовка данных и предварительный анализ	8
2.2. Линейные методы классификации	8
2.3. Сравнительный анализ моделей	8
2.4. Кривые обучения моделей	9
2.5. Визуализация разделяющих поверхностей	10
3. Заключение	13

1. Введение

Классификация является одной из фундаментальных задач машинного обучения, направленной на автоматическое отнесение объектов к одному из заранее определённых классов на основе их признаков. Применения включают медицину, распознавание образов, биоинформатику и финансы.

Целью данной работы является исследование и сравнение методов классификации: линейных (логистическая регрессия, линейный SVM), нелинейных (SVM с RBF-ядром), байесовских (Gaussian Naive Bayes) и методов дискриминантного анализа (LDA, QDA) на примере датасета Breast Cancer Wisconsin. Датасет содержит $n = 569$ наблюдений, $p = 30$ непрерывных признаков и два класса (доброкачественные / злокачественные), причём классы частично несбалансированы.

Основные модели и используемые формулы

1. Логистическая регрессия. Модель предсказывает апостериорную вероятность класса через сигмоиду:

$$P(y_i = 1 \mid x_i) = \sigma(z_i) = \frac{1}{1 + \exp(-z_i)}, \quad z_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}. \quad (1)$$

Функция потерь (логистический лог-loss, отрицательное лог-правдоподобие):

$$L_{\log}(\beta) = - \sum_{i=1}^n (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)), \quad (2)$$

где $\hat{y}_i = P(y_i = 1 \mid x_i)$.

2. Линейный SVM. Решение задаётся гиперплоскостью $f(x) = w^\top x + b$. Класс определяется знаком:

$$\hat{y} = \text{sign}(w^\top x + b). \quad (3)$$

Стандартная оптимизационная задача (с мягкими ограничениями) и hinge-loss:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i, \quad (4)$$

$$\text{при } y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad (5)$$

или эквивалентно минимизации суммы hinge-потерь

$$L_{hinge} = \sum_{i=1}^n \max(0, 1 - y_i(w^\top x_i + b)). \quad (6)$$

3. SVM с RBF-ядром. Нелинейный SVM использует ядро $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ и работает в пространстве $\phi(x)$; оптимизация аналогична, но с kernel trick.

4. Наивный байесовский классификатор (Gaussian NB). По теореме Байеса:

$$P(C_k | x) = \frac{P(x | C_k)P(C_k)}{P(x)}. \quad (7)$$

При предположении независимости признаков и нормальности:

$$P(x | C_k) = \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\left(-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right). \quad (8)$$

5. Дискриминантный анализ (LDA/QDA). LDA предполагает общую ковариацию Σ :

$$\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log \pi_k, \quad (9)$$

QDA использует отдельные ковариации Σ_k :

$$\delta_k^Q(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k) + \log \pi_k. \quad (10)$$

Метрики качества и их формулы

Для объективного сравнения моделей используются следующие метрики, вычисляемые по матрице ошибок

$$\begin{pmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{pmatrix}.$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (11)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (12)$$

$$\text{Recall (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (13)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (14)$$

Также вводят FPR (false positive rate):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (15)$$

При несбалансированных классах полезны макро- и взвешенные усреднения метрик:

- macro-average — среднее показателей по классам (без учёта размера класса),
- weighted-average — среднее с весами, пропорциональными support (количеству объектов класса).

Кривые обучения и кросс-валидация

Кривая обучения строится по значениям точности на обучающей и валидационной выборках при различном размере обучающей подвыборки m :

$$\text{score}_{\text{train}}(m), \quad \text{score}_{\text{cv}}(m).$$

На практике для набора размеров $\{m_1, \dots, m_k\}$ вычисляем средние значения по фолдам CV:

$$\overline{\text{score}}_{cv}(m) = \frac{1}{K} \sum_{r=1}^K \text{score}^{(r)}(m).$$

Кросс-валидация оценивает стабильность модели:

$$\text{CV_score} = \frac{1}{K} \sum_{r=1}^K \text{metric}(\text{model trained on fold } r). \quad (16)$$

Примечания к реализации

- Перед обучением для SVM и логистической регрессии применяется масштабирование признаков (StandardScaler).
- В работе оцениваются время обучения и время предсказания моделей, поскольку в практических задачах производительность имеет значение.
- Для визуализации поведения моделей используются матрицы ошибок, кривые обучения и 2D-плоты разделяющих поверхностей (для LDA/QDA на выбранной паре признаков).

2. Основная часть

2.1. Подготовка данных и предварительный анализ

В данной работе используется датасет Breast Cancer Wisconsin Dataset.

1. Загрузка и предобработка данных: удаление пропусков, проверка типов признаков.
2. Анализ распределения классов. Распределение классов показано на Рис. ??.
3. Исследование корреляций между признаками. Корреляционная матрица представлена на Рис. 2.
4. Разделение данных на обучающую и тестовую выборки в соотношении 80/20.
5. Масштабирование признаков для методов, требующих нормализации (SVM, логистическая регрессия).

Описание выбранного датасета

- Название датасета: Breast Cancer Wisconsin Dataset
- Описание: Классификация опухолей молочной железы на доброкачественные и злокачественные на основе признаков, полученных с помощью цифровой обработки изображения.
- Количество наблюдений: 569
- Количество классов: 2
- Количество признаков: 30
- Тип признаков: непрерывные
- Проблема: несбалансированная классификация (357 доброкачественных, 212 злокачественных)

Рисунок 1 – Распределение классов

2.2. Линейные методы классификации

2.3. Сравнительный анализ моделей

Сравнение всех моделей по метрикам качества, времени обучения и предсказания представлено на Рис. 4.

В данном разделе показаны все ключевые этапы эксперимента:

- Подготовка и анализ данных.
- Построение линейных, нелинейных и байесовских моделей.
- Применение методов дискриминантного анализа.
- Визуализация результатов в виде матриц ошибок и сравнительных таблиц.

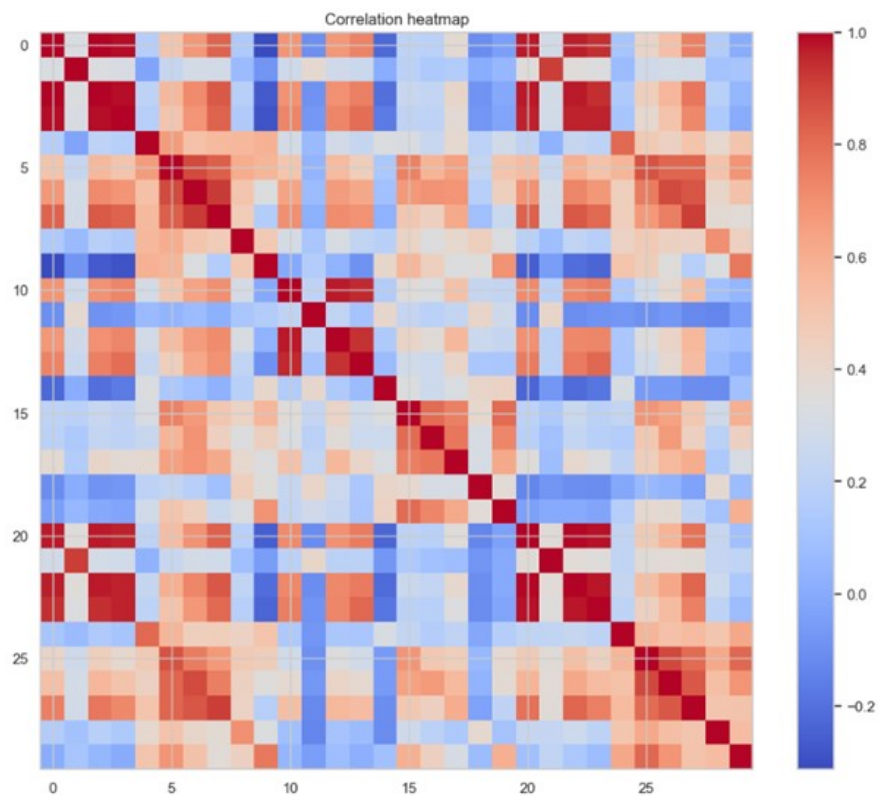


Рисунок 2 – Корреляционная матрица признаков

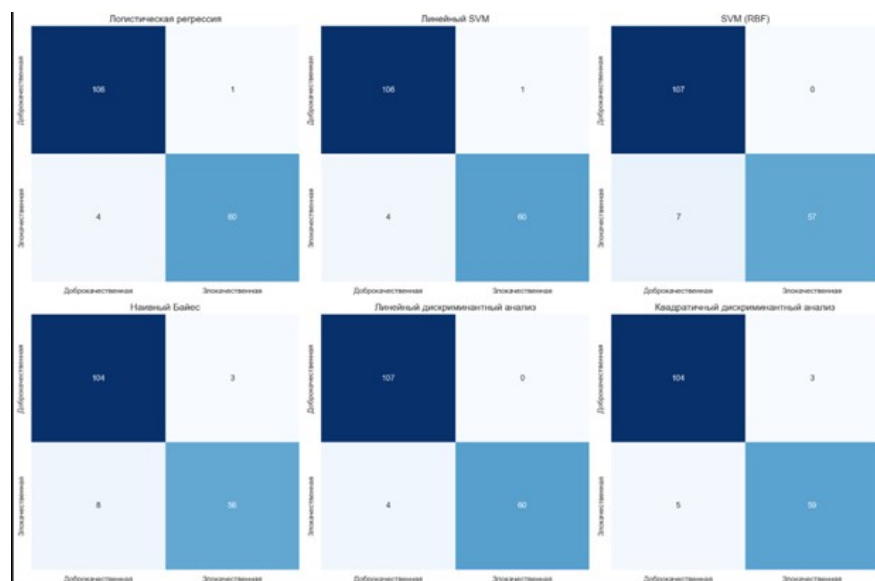


Рисунок 3 – Матрицы

2.4. Кривые обучения моделей

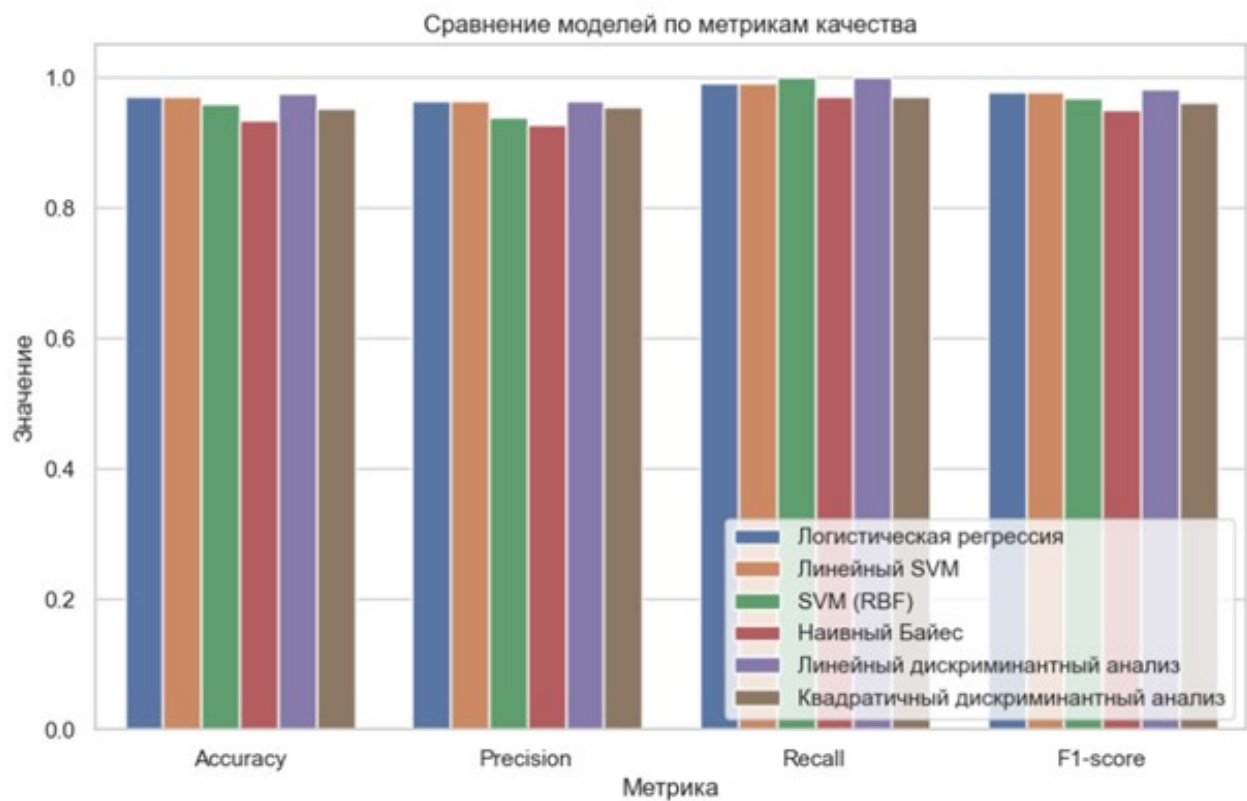


Рисунок 4 – Сравнение моделей по метрикам качества

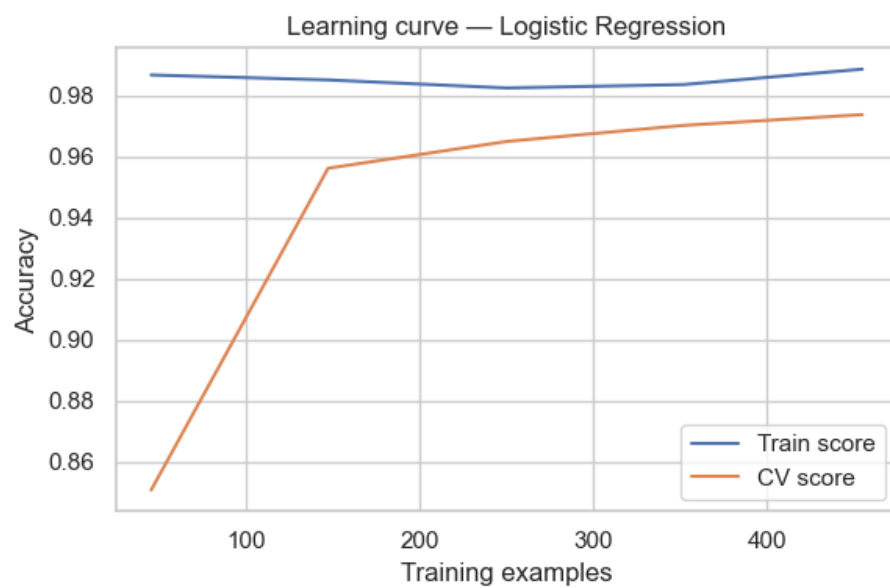


Рисунок 5 – Кривая обучения логистической регрессии

2.5. Визуализация разделяющих поверхностей

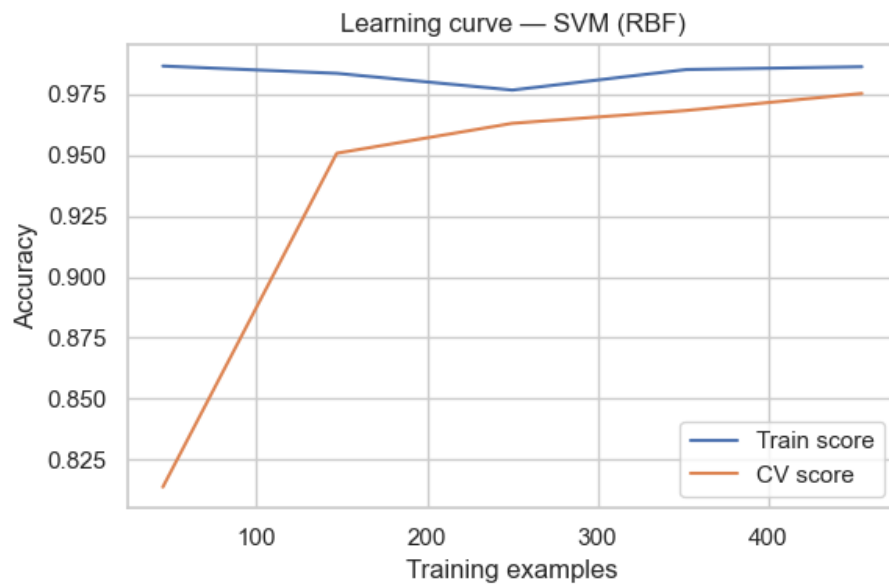


Рисунок 6 – Кривая обучения SVM с RBF-ядром

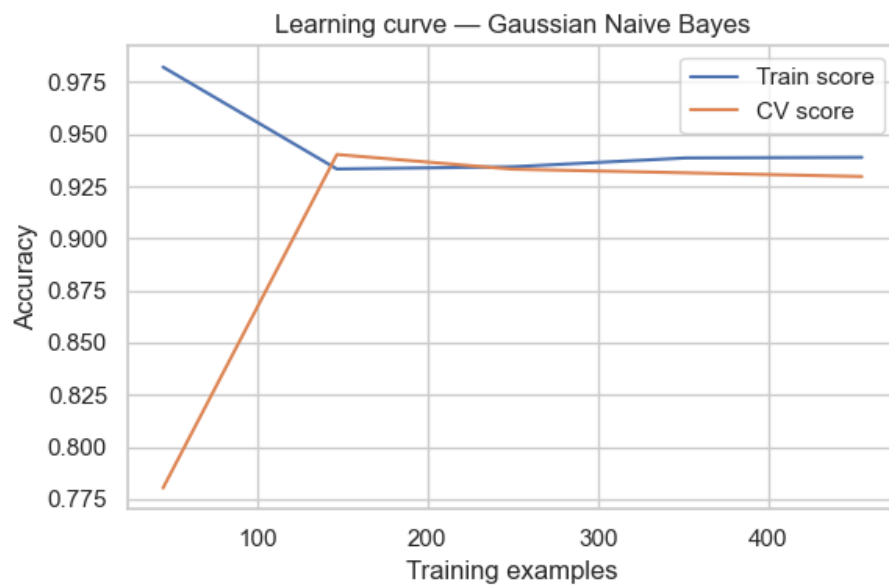


Рисунок 7 – Кривая обучения наивного байесовского классификатора (Gaussian NB)

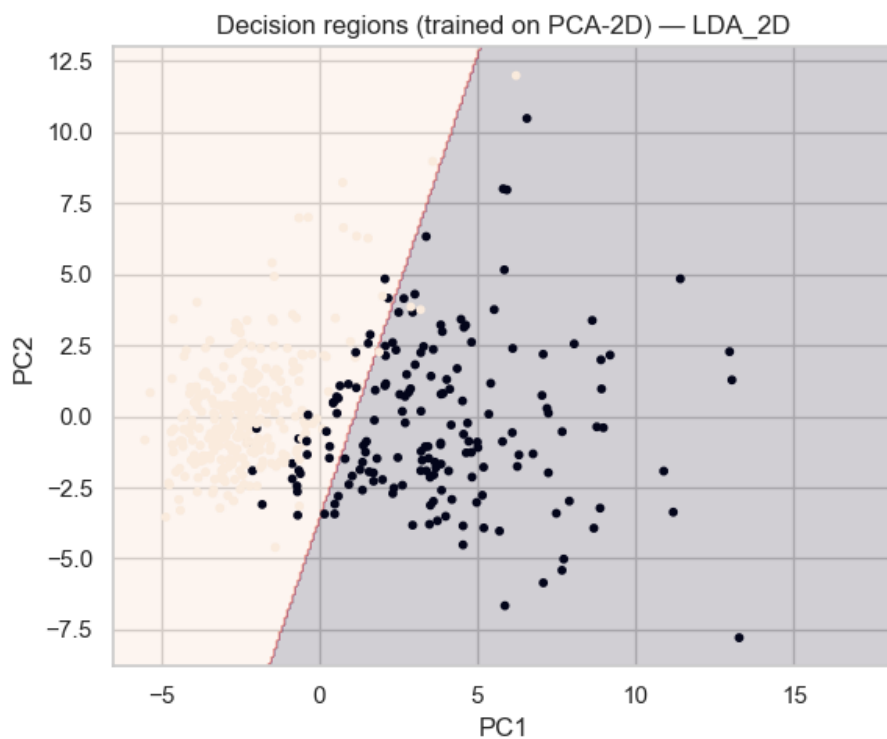


Рисунок 8 – Разделяющая поверхность модели LDA в двухмерном пространстве признаков

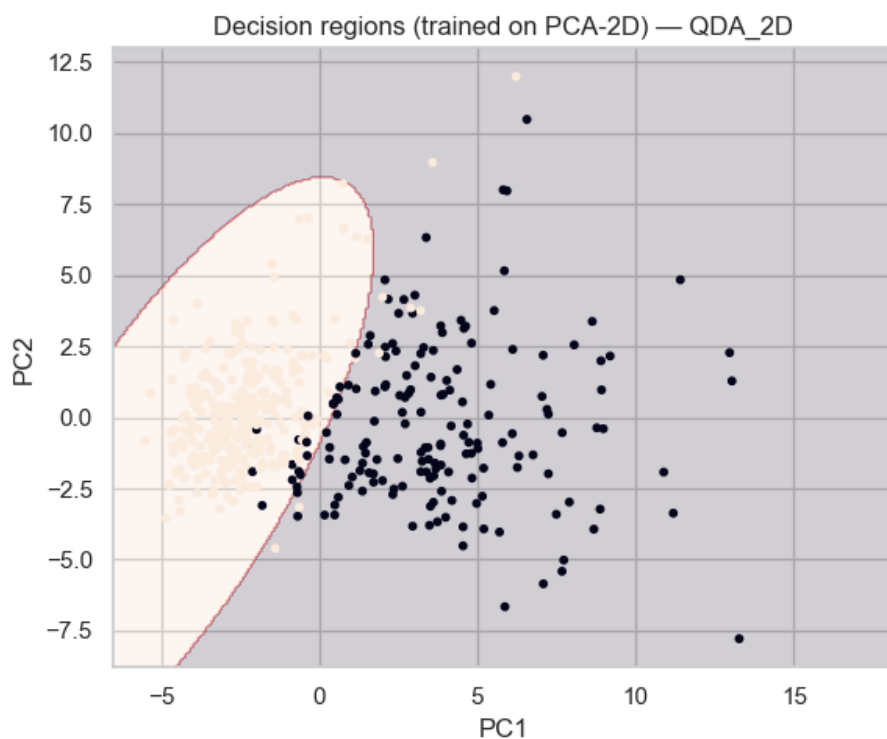


Рисунок 9 – Разделяющая поверхность модели QDA в двухмерном пространстве признаков

3. Заключение

В данной работе были исследованы и сравнены различные методы классификации на основе датасета Breast Cancer Wisconsin (Diagnostic), широко применяемого для задач медицинской диагностики. Цель исследования заключалась в построении нескольких моделей классификации, оценке их качества и сравнении линейных, нелинейных и байесовских подходов, а также методов дискриминантного анализа.

В рамках эксперимента были реализованы следующие методы: логистическая регрессия, линейный SVM, SVM с RBF-ядром, наивный байесовский классификатор, модели LDA и QDA. Для оценки качества использовались метрики accuracy, precision, recall, F1-score, а также анализировались матрицы ошибок.

Полученные результаты позволяют сформулировать следующие выводы:

- Логистическая регрессия и линейный SVM показали идентичные результаты (accuracy = 0.97), высокую точность и чувствительность. Модели почти не допускают ошибок на классе «доброкачественные опухоли» и демонстрируют устойчивое поведение.
- SVM с RBF-ядром также показал высокое качество (accuracy = 0.96), однако модель сильнее ошибается на классе «злокачественные». Она идеально классифицирует доброкачественные опухоли (recall = 1.00), но имеет больше ложных отрицаний.
- Наивный Байес продемонстрировал худшее качество среди рассмотренных моделей (accuracy = 0.94). Это связано с тем, что допущение о независимости признаков плохо выполняется в данном наборе данных.
- Линейный дискриминантный анализ (LDA) показал наилучший результат среди всех моделей: accuracy = 0.9766, идеальная чувствительность по доброкачественному классу и высокая точность по злокачественному. Это указывает на то, что классы в данных достаточно хорошо разделимы линейной границей.
- Квадратичный дискриминантный анализ (QDA) продемонстрировал более низкое качество по сравнению с LDA (accuracy = 0.9532), что объясняется сильной чувствительностью QDA к выбросам и малому числу наблюдений в некоторых группах.

Сравнение времени обучения и предсказания показало, что:

- самые быстрые модели — Наивный Байес и линейный SVM;
- SVM (RBF) требует больше времени на предсказание, что связано с вычислением ядра;
- LDA обеспечивает лучший баланс между скоростью и качеством.

На основании всех метрик можно заключить, что наилучшим алгоритмом для данного набора данных является линейный дискриминантный анализ (LDA), так как он демонстрирует максимальное значение ассурасу, F1-меры и стабильную работу на обоих классах.

В целом проведённое исследование подтверждает, что:

1. линейные методы классификации являются эффективными для медицинских данных, обладающих чёткой структурой и разделимостью;
2. SVM с RBF-ядром целесообразно использовать в задачах, где классы разделены нелинейно, однако на данном датасете его преимущества выражены слабее;
3. байесовские методы подходят для быстрого базового анализа, но хуже справляются с сильно коррелированными признаками;
4. дискриминантный анализ может превосходить даже современные методы, если выполнены его предпосылки.

Таким образом, выполненная работа демонстрирует важность сравнительного анализа нескольких методов машинного обучения и подтверждает, что качество классификации существенно зависит от структуры данных и от соответствия модели их статистическим свойствам.

Список литературы

- [1] Scikit-Learn: Machine Learning in Python. URL: <https://scikit-learn.org/>
- [2] Pandas Documentation. URL: <https://pandas.pydata.org/>
- [3] Statsmodels: Statistical Models in Python. URL: <https://www.statsmodels.org/>
- [4] Matplotlib Python plotting library. URL: <https://matplotlib.org/>
- [5] UCI Machine Learning Repository – California Housing dataset. URL: <https://archive.ics.uci.edu/>