Лабораторная работа №4. Простейшие методы классификации данных

Основное задание (**4 балла**)

1. Загрузить данные согласно своему варианту. Данные представляют собой таблицу, состоящую из 5 столбцов: 1 столбец – это номер класса, к которому принадлежит наблюдение, 2-4 столбцы – некоторые переменные, которыми задано текущее наблюдение.

2. Удалить из выборки 15 случайных наблюдений (сохранить их отдельно). Оставшуюся выборку назовем обучающей. По ней мы будет обучать классификаторы. Сохраненные отдельно 15 наблюдений будем называть проверочной выборкой, по ней мы будем проверять качество классификатора.

3. Построить линейный классификатор для решения задачи классификации (см. файл Filzmoser-Lections.pdf стр. 50-51). Метод основан на применении модели линейной регрессии для классификации. Реализовать классификатор означает обучить его на имеющейся выборке (если это необходимо), после чего уметь классифицировать любое новое наблюдение. Обучить классификатор на обучающей выборке. Применить его к проверочной выборке. Вывести количество ошибок, которое дал построенный классификатор для проверочной выборки.

4. Реализовать метод классификации k-ближайших соседей (ссылка, kNN, k-nearest neighbors). Для $k = 3$ и $k = 5$. Для нахождения расстояния использовать классическую Евклидову метрику. Применить его к проверочной выборке. Вывести количество ошибок, которое дал построенный классификатор для проверочной выборки.

Бонусные задания

Графическая иллюстрация работы классификатора.

5. (**0,5 балла**) Построить диаграмму рассеяния некоторых 2 переменных из переменных 2-4. Цвет и форму точек менять в зависимости от номера класса.

6. (**1 балл**) Построить линейный классификатор, на основе только 2 выбранных в пункте 5 переменных. Изобразить на этой диаграмме границы классов, получаемые для построенного линейного классификатора.

7. (**0,5 балла**) Построить границы остальных классификаторов, построенных на основе выбранных 2 переменных.

Другие методы классификации

8. **(1 балл)** Реализовать метод линейного дискриминантного анализа (см. файл Filzmoser-Lections.pdf стр. 51-53). Обучить классификатор на обучающей выборке. Применить его к проверочной выборке. Вывести количество ошибок, которое дал построенный классификатор для проверочной выборки.

9. **(1 балл)** Реализовать метод квадратичного дискриминантного анализа (см. файл Filzmoser-Lections.pdf стр. 53). Обучить классификатор на обучающей выборке. Применить его к проверочной выборке. Вывести количество ошибок, которое дал построенный классификатор для проверочной выборки.

10. **(1 балл)** Разобраться в методе опорных векторов (svm, support vector machine) для 2 классов. В выборке оставить наблюдения только из 2 классов. Применить готовую реализацию. Проинтерпретировать полученные результаты. Проиллюстрировать результаты работы метода.

Перекрёстная проверка (кросс-валиадция, cross-validation)

Часто на практике количество размеченных данных ограничено (данные, для которых для наблюдений известны номера классов). В этом случае применить Метод-Карло напрямую не получается: его мы применяли в прошлой лабораторной работе, когда для оценки среднеквадратической ошибки, мы генерировали выборку много раз и усредняли квадрат ошибки прогнозирования. Однако, сравнивать разные методы классификации хотелось бы. Поэтому поступают следующим образом:

Фиксируем некоторую долю, например, $p = 10\%$. (*) Затем случайным образом выбираем p-ую часть наблюдений и формируем из них тестовую выборку, из остальных наблюдений формируем обучающую выборку. Проводим обучение сравниваемых моделей на обучающей выборке и

вычисляем вероятности ошибочных классификаций по тестовой выборке. Повторяем все операции, начиная с (*), M раз (например, $M = 100$) и усредняем полученные вероятности ошибок. Тот метод, который дал наименьшие вероятности ошибок, считается предпочтительным для конкретной задачи.

11. **(1 балл)** Для всех реализованных методов классификации данных реализовать описанный выше метод сравнения.

Уменьшение числа переменных в модели*

12. (**2 балла**) На практике часто бывает так, что использование всех имеющихся переменных только ухудшает качество итоговой модели (метода). Часто имеет смысл удалить из модели часть переменных, в этом случае качество классификации (прогноза и т.п.) может значительно улучшиться. Придумать (найти в литературе) подходы по удалению «лишних» переменных. Оставить лишь 2 переменных в модели. Пункты 5-7 реализовать для выбранных двух переменных.