Лабораторная работа №7  
**«Классификация с использованием Наивного байесовского классификатора».**

**Часть 1.** Рассмотрите задачу классификации на классическом примере набора «Ирисы» (**4 балла**).

Необходимо построить (реализовать на языке Python) наивный байесовский классификатор на основе, обучив его на данных обучающей выборки в файле «iris\_data\_Var.txt» (для своего варианта, номер варианта – порядковый номер студента). Полагать, что признаки независимы и распределены по гауссовскому закону (нормальной плотности распределения вероятностей).

Выделить обучающую выборку (70% всей выборки, случайное разделение реализовать самостоятельно) и отобразить в виде графика точек на плоскости (объекты разных классов должны быть иметь разные маркеры и цвет, сделавшим в классе +1 балл). Обучить модель. Отобразить тестовую выборку, чтобы было понятно качество классификации объектов из неё..

Выполнить кросс-валидацию (реализовать самостоятельно). Оценить качество модели.

**Часть 2.** Самостоятельно реализовать наивный байес для задачи обнаружения спама (на основе данных из файла «SpamDetectionData.txt», **6 баллов**). В качестве метрики для сравнения использовать точность (accuracy) – процент правильно классифицированных примеров из тестовой выборки. Сравнить точность на train/ test частях.

1. Загрузить обучающую выборку в 2 списка – позитивные и негативные отзывы. Чему равна минимальная, максимальная, средняя, медианная длина (в символах) позитивных / негативных отзывов? Показать диаграмму долей разных классов отзывов в наборе.
2. Сделать предобработку. Перевести отзывы в нижний регистр. Подумать, как быть со знаками препинания, цифрами, иными символами.
3. Сделать токенизацию – то есть представить каждый отзыв в виде списка токенов (слов, чисел, знаков пунктуации). Каждый токен станет элементом словаря. При токенизации учитывайте, что в текстах знаки пунктуации приклеиваются к предшествующему слову (не отделяются пробелом). Предобработку и токенизацию полезно выделить в отдельные функции, т.к. они понадобятся и при обработке тестовой выборки. Тестовая и обучающая выборка должны обрабатываться одинаково.
4. Построить 2 словаря {слово:частота} с частотами каждого слов в позитивных и негативных отзывах. Для оботх словарей распечатайте 20 слов с максимальными и 20 слов с минимальными наивными байесовскими весами (наивным байесовским весом слова будем называть log [P(w|pos) / P(w|neg)]), сами веса, а также абсолютные частоты этих слов в позитивном и негативном классах. Результаты приведите в виде таблицы. Какие слова из полученных списков кажутся позитивно/негативно окрашенными, но имеют низкий вес?
5. Используя формулы из теории реализовать байесовские классификаторы. Обучить классификаторы на обучающей выборке и разметить с их помощью тестовую выборку. Сколько времени обучается классификаторы и сколько времени уходит на обработку тестовой выборки? Чему равна точность классификаторов на обучающей и тестовой выборках? Результаты приведите в виде таблицы. Как вы думаете, с чем связана разная точность на разных выборках?

По результатам работы подготовить отчёт в соответствии с требованиями к отчётам предыдущих работ.

Программа защищается на паре.

Можно работать в группе по 2 человека.