**Лабораторная работа №6 Метод опорных векторов**

**PLA vs. SVM**

В следующих проблемах мы сравниваем PLA с SVM (используйте estimator SVC в scikit-learn) с жестким отступом на линейно разделяемых наборах данных. Для каждого прогона вы создадите свою целевую функцию f и набор данных D. Возьмите d = 2 и выберите случайную линию на плоскости в качестве целевой функции f (сделайте это, взяв две случайные равномерно распределенные точки на [-1; 1]x [-1; 1] проведите линию через них), где одна сторона линии отображает +1, а остальные точки - -1. Выберите xn набора данных как случайные точки

в X = [-1; 1]x [-1; 1] и оцените целевую функцию на каждом xn, чтобы получить соответствующий выход yn. Если все точки данных находятся на одной стороне линии, то запустите новый прогон.

Начните PLA с вектора с нулевыми значениями и выберите случайным образом точку с неправильной оценкой для каждой итерации PLA. Запустите PLA для определения гипотезы g и измерьте несогласие между f и gPLA как

P [f (x) ≠ gPLA (x)] (вы можете либо вычислить это точно или приблизительно, создав достаточно большой отдельный набор точек для проверки).

Теперь запустите SVM по тем же данным, чтобы найти финальную гипотезу gSVM. Измерьте несогласованность между f и gSVM как P [f (x) ≠ gSVM (x)] и подсчитайте количество поддерживающих векторов, которые вы получаете в каждом прогоне.

1. При N = 10 повторите описанный выше эксперимент 1000 раз. Как часто gSVM лучше gPLA в приближении f? Процент близок к:

[a] 20%

[b] 40%

[c] 60%

[d] 80%

[e] 100%

1. При N = 100 повторите описанный выше эксперимент 1000 раз. Как часто gSVM лучше gPLA в приближении f? Процент близок к:

[a] 10%

[b] 30%

[c] 50%

[d] 70%

[e] 90%

3. Для N = 100 наиболее близкое к среднему числу поддерживающих векторов gSVM (усредненный по 1000 прогонам)?

[a] 2

[b] 3

[c] 5

[d] 10

[e] 20

**SVM с мягким отступом**

Сейчас применим SVM с мягким отступом к рукописным цифрам из обработанного набора данных почтовых индексов US Postal Service. Загрузите данные (извлеченные признаки: интенсивность и симметрия) для обучения и тестирования:

http://www.amlbook.com/data/zip/features.train

<http://www.amlbook.com/data/zip/features.test>

(формат каждой строки: **digit intensity symmetry**). Мы будем обучать два типа двоичных классификаторов: «один- против –одного» (одна цифра - класс +1, а другая цифра - класс -1, а остальные цифры не учитываются) и «один-против-всех» (одна цифра - это класс + 1, а остальные цифры – класс 1).

Примените SVM с мягким отступом по вышеуказанному набору данных zip-кода.

**Полиномиальное ядро**

Будем рассматривать полиномиальное ядро 

Гдн Q – степень полинома

1. Пусть С=0.01, Q=2, какой из классификаторов имеет наибольшую ошибку Ein?

[a] 0 versus all

[b] 2 versus all

[c] 4 versus all

[d] 6 versus all

[e] 8 versus all

1. Пусть С=0.01, Q=2, какой из классификаторов имеет наименьшую ошибку Ein?

[a] 1 versus all

[b] 3 versus all

[c] 5 versus all

[d] 7 versus all

[e] 9 versus all

6. Рассмотрим классификатор «1 против 5» с Q = 2 и C={0. 001; 0.01; 0. 1; 1}.

Какое из следующих утверждений верно?

[a] Число поддерживающих векторов уменьшается, когда C растет.

[b] Число поддерживающих векторов возрастает, когда C растет.

[c] Eout падает, когда C растет.

[d] Максимум C дает самое низкое Ein.

[e] Ничего из вышеперечисленного

7. В классификатор «1 против 5», сравнивая Q = 2 и Q = 5, какое из следующих утверждений верно?

[a] Когда C = 0,0001, Ein выше при Q = 5.

[b] Когда C = 0,001, число опорных векторов ниже при Q = 5.

[c] Когда C = 0.01, Ein выше при Q = 5.

[d] Когда C = 1, Eout ниже при Q = 5.

[e] Ничего из вышеперечисленного

**Кросс-валидация**

В следующих двух проблемах мы будем экспериментировать с 10-кратной перекрестной проверкой для полиномиального ядра. Поскольку Ecv - это случайная величина, зависящая от случайного разбиения данных, мы попробуем 100 прогонов с разным разбиением и основываем наш ответ на том, сколько пробегов приводит к определенному выбору.

8. Рассмотрим классификатор «1 против 5» с Q = 2. Мы используем Ecv для выбора C из{0.0001; 0.001; 0.01; 0.1; 1}. Если в Ecv есть одинаковые, выберите меньшее значение C. В пределах 100 случайных прогонов, какое из следующих утверждений верно?

[a] C = 0. 0001 выбирается наиболее часто.

[b] C = 0.001 выбирается наиболее часто.

[c] Наиболее часто выбирается C = 0.01.

[d] Наиболее часто выбирается C = 0.1.

[e] C = 1 выбирается наиболее часто.

9.Снова рассмотрим классификатор «1 против 5» с Q = 2. Для модели выбранной в п.8, среднее значение ошибки кросс-валидации Ecv за 100 прогонов ближе всего:

(a) 0:001

(b) 0:003

(c) 0:005

(d) 0:007

(e) 0:009

**RBF ядро**

Рассмотрим ядро радиальной базисной функции (RBF) K(xn; xm)=exp(-||xn- xm||2) в SVM с мягким запасом. Рассмотрим классификатор «1 против 5».

10. Какое из следующих значений C приводит к наименьшему Ein?

[a] C = 0.01

[b] C = 1

[c] C = 100

[d] C = 104

[e] C = 106

11. Какое из следующих значений C приводит к наименьшему Eout?

[a] C = 0.01

[b] C = 1

[c] C = 100

[d] C = 104

[e] C = 106