Министерство образования Республики Беларусь

БЕЛОРУССКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ

Кафедра «Программное обеспечение информационных систем и технологий»

Методические указания

к лабораторной работе по теме

**«Простейшие методы классификации»**

для студентов специальностей

1- 40 01 01 «Программное обеспечение информационных технологий» и

1-40 05 01 «Информационные системы и технологии (в проектировании и производстве)»

|  |  |
| --- | --- |
| Разработчик  к.т.н., доцент кафедры ПОИСиТ | Ковалева И.Л. |

Минск

БНТУ

2022

**ЛАБОРАТОРНАЯ РАБОТА**

**Простейшие методы классификации**

Цель работы: изучение наиболее простых методов классификации.

**Теоретическая часть**

Классификация – это отнесение конкретного объекта (реализа­ции), представленного значениями его свойств (признаков), к одному из фиксированного перечня образов (классов) по определённому решающему правилу (алгоритму) в соответствии с поставленной целью.

Отсюда следует, что классификация может осуществляться любой системой (живой или неживой), выполняющей следующие функции: измерение значений признаков, производство вычислений, реализующих решающее правило. При этом перечень образов, информативных признаков и решающие правила либо задаются системе классификации извне, либо формируются самой системой. Вспомогательная, но важная функция систем классификации – оценка риска потерь. Без этой функции невозможно, например, построить оптимальные решающие правила, выбрать наиболее информативную систему признаков, которые используются при классификации, и др. Рассмотрим простейшие методы классификации.

1. **Метод «k-ближайших соседей»**

В зависимости от числа исследуемых соседей принято выделять следующие варианты метода:

**Метод ближайшего соседа** является, пожалуй, самым простым [алгоритмом](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC) [классификации](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F). Классифицируемый объект x относится к тому классу y_i, которому принадлежит ближайший объект обучающей выборки x_i.

**Метод k ближайших соседей**. Для повышения надёжности классификации объект относится к тому классу, которому принадлежит большинство из его *соседей* — k ближайших к нему объектов обучающей выборки x_i. В задачах с двумя классами число соседей берут нечётным, чтобы не возникало ситуаций неоднозначности, когда одинаковое число соседей принадлежат разным классам.

**Метод взвешенных ближайших соседей**. В задачах с числом классов 3 и более нечётность уже не помогает, и ситуации неоднозначности всё равно могут возникать. Тогда i-му соседу приписывается вес w_i, как правило, убывающий с ростом ранга соседа i. Объект относится к тому классу, который набирает больший суммарный вес среди k ближайших соседей.

**Выбор числа соседей k**

При k=1 [алгоритм](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC) ближайшего соседа неустойчив к шумовым выбросам: он даёт ошибочные классификации не только на самих объектах-выбросах, но и на ближайших к ним объектах других классов. При k=m, наоборот, [алгоритм](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC) чрезмерно устойчив и вырождается в константу. Таким образом, крайние значения k нежелательны. На практике оптимальное значение параметра k определяют по критерию [скользящего контроля](http://www.machinelearning.ru/wiki/index.php?title=%D0%A1%D0%BA%D0%BE%D0%BB%D1%8C%D0%B7%D1%8F%D1%89%D0%B8%D0%B9_%D0%BA%D0%BE%D0%BD%D1%82%D1%80%D0%BE%D0%BB%D1%8C), чаще всего — методом исключения объектов по одному (leave-one-out cross-validation).

**Отсев шума (выбросов)**

Обычно объекты обучения не являются равноценными. Среди них могут находиться типичные представители классов — *эталоны*. Если классифицируемый объект близок к эталону, то, скорее всего, он принадлежит тому же классу. Ещё одна категория объектов — *неинформативные* или *периферийные*. Они плотно окружены другими объектами того же класса. Если их удалить из выборки, это практически не отразится на качестве классификации. Наконец, в выборку может попасть некоторое количество *шумовых выбросов* — объектов, находящихся «в гуще» чужого класса. Как правило, их удаление только улучшает качество классификации.

Исключение из выборки шумовых и неинформативных объектов даёт несколько преимуществ одновременно: повышается качество классификации, сокращается объём хранимых данных и уменьшается время классификации, затрачиваемое на поиск ближайших эталонов.

**Проблема выбора метрики**

Это наиболее сложная из всех проблем. В практических задачах классификации редко встречаются такие «идеальные случаи», когда заранее известна хорошая функция расстояния \rho(x,x'). Если объекты описываются числовыми векторами, часто берут евклидову метрику. Этот выбор, как правило, ничем не обоснован — просто это первое, что приходит в голову. При этом необходимо помнить, что все признаки должны быть измерены «в одном масштабе», а лучше всего — отнормированы. В противном случае признак с наибольшими числовыми значениями будет доминировать в метрике, остальные признаки, фактически, учитываться не будут.

Однако и нормировка является весьма сомнительной эвристикой, так как остаётся вопрос: «неужели все признаки одинаково значимы и должны учитываться примерно с одинаковым весом?»

Если признаков слишком много, а расстояние вычисляется как сумма отклонений по отдельным признакам, то возникает [проблема проклятия размерности](http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D1%80%D0%BE%D0%BA%D0%BB%D1%8F%D1%82%D0%B8%D0%B5_%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D0%B8). Суммы большого числа отклонений с большой вероятностью имеют очень близкие значения (согласно [закону больших чисел](http://www.machinelearning.ru/wiki/index.php?title=%D0%97%D0%B0%D0%BA%D0%BE%D0%BD_%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D1%85_%D1%87%D0%B8%D1%81%D0%B5%D0%BB)). Получается, что в пространстве высокой размерности все объекты примерно одинаково далеки друг от друга; выбор k ближайших соседей становится практически произвольным.

Проблема решается путём отбора относительно небольшого числа [информативных признаков](http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D1%82%D0%B1%D0%BE%D1%80_%D0%BF%D1%80%D0%B8%D0%B7%D0%BD%D0%B0%D0%BA%D0%BE%D0%B2&action=edit) (features selection). В [алгоритмах вычисления оценок](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC_%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F_%D0%BE%D1%86%D0%B5%D0%BD%D0%BE%D0%BA) строится множество различных наборов признаков (т.н. опорных множеств), для каждого строится своя функция близости, затем по всем функциям близости производится голосование.

**2.Метод "Близость к эталону класса"**

Для каждого класса по обучающей выборке строится эталон, имеющий значения признаков

,

где =,

 – количество объектов данного образа в обучающей выборке,

 – номер признака. 

По существу, эталон – это усреднённый по обучающей выборке абстрактный объект (рисунок 1). Абстрактным мы его называем потому, что он может не совпадать не только ни с одним объектом обучающей выборки, но и ни с одним объектом генеральной совокупности.

Классификация осуществляется следующим образом. На вход системы поступает объект , принадлежность которого к тому или иному образу системе неизвестна. От этого объекта измеряются расстояния до эталонов всех образов, и  система относит к тому образу, расстояние до эталона которого минимально. Расстояние измеряется в той метрике, которая введена для решения определённой задачи распознавания.

× ×

× × ×

× × × ×

× × ×

× ×

×

×

Рисунок 1- Решающее правило "Минимум расстояния   
до эталона класса":

 – эталон первого класса,

×

 – эталон второго класса

**3.Метод потенциальных функций**

Название метода в определённой степени связано со следующей аналогией (для простоты будем считать, что распознаётся два образа). Представим себе, что объекты являются точками некоторого пространства . В эти точки будем помещать заряды , если объект принадлежит образу , и , если объект принадлежит образу  (рисунок 2).



Рисунок 2 - Иллюстрация синтеза потенциальной функции   
в процессе обучения:

 – потенциальная функция, порождаемая одиночным объектом;

 – суммарная потенциальная функция, порождённая обучающей последовательностью

Функцию, описывающую распределение электростатического потенциала в таком поле, можно использовать в качестве решающего правила (или для его построения). Если потенциал точки , создаваемый единичным зарядом, находящимся в , равен , то общий потенциал в , создаваемый  зарядами, равен



 – потенциальная функция. Она, как в физике, убывает с ростом евклидова расстояния между  и . Чаще всего в качестве потенциальной используется функция, имеющая максимум при  и монотонно убывающая до нуля при .

Классификация может осуществляться следующим способом. В точке , где находится неопознанный объект, вычисляется потенциал . Если он оказывается положительным, то объект относят к образу . Если отрицательным – к образу .

При большом объёме обучающей выборки эти вычисления достаточно громоздки, и зачастую выгоднее вычислять не , а оценивать разделяющую классы (образы) границу либо аппроксимировать потенциальное поле.

Выбор вида потенциальных функций – дело непростое. Например, если они очень быстро убывают с ростом расстояния, то можно добиться безошибочного разделения обучающих выборок. Однако при этом возникают определённые неприятности при распознавании неопознанных объектов (снижается достоверность принимаемого решения, возрастает зона неопределённости). При слишком "пологих" потенциальных функциях может необоснованно увеличиться количество ошибок распознавания, в том числе и на обучающих объектах. Определённые рекомендации в этом отношении можно получить, рассматривая метод потенциальных функций со статистических позиций (восстановление плотности распределения вероятностей  или разделяющей границы по выборке с использованием процедуры типа стохастической аппроксимации).

**Гипотеза компактности**

Все перечисленные методы неявно опираются на одно важное предположение, называемое [гипотезой компактности](http://www.machinelearning.ru/wiki/index.php?title=%D0%93%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0_%D0%BA%D0%BE%D0%BC%D0%BF%D0%B0%D0%BA%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B8): *если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных*. В этом случае граница между классами имеет достаточно простую форму, а классы образуют компактно локализованные области в пространстве объектов. (Заметим, что в математическом анализе *компактными* называются ограниченные замкнутые множества. *Гипотеза компактности* не имеет ничего общего с этим понятием, и должна пониматься скорее в «бытовом» смысле этого слова.)

**Задание по лабораторной работе**

1. Изучить и адаптировать заданный преподавателем метод классификации для классов символов, выбранных в первой лабораторной работе. В качестве признаков, описывающих объекты классификации, использовать количество узловых и концевых точек (рисунок 3).

******

Рисунок 3 – Признаки для описания простых изображений

1. Описание объектов для каждого способа представить в виде таблиц Excel.
2. Разработать программу, выполняющую классификацию объектов методом, заданным преподавателем.
3. Результаты расчетов необходимых расстояний внести в таблицы Excel.
4. Выполнить визуализацию результатов классификации для 1 способа описания (через количество узловых и концевых точек)
5. Оформить отчет, описав все разработанные признаки.

**Пример визуализация результатов классификации методом «близость к эталонам»**

|  |  |
| --- | --- |
| 1. **Вычисление эталонов и расстояний от неизвестной точки до эталонов всех классов** | 1. **Визуализация** |