0.1 Основные понятия

1. **Задача классификации** сводится к определению класса объекта по его характеристикам. Множество классов известно заранее.
   Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных.
   Задача регрессии подобно задаче классификации позволяет определить по известным характеристикам объекта значение некоторого параметра из множества действительных чисел.
   Supervised leraning задача анализа данных решается поэтапно. Сначала строится модель анализируемых данных – классификатор. Затем классификатор подвергается обучению. Другими словами, проверяется качество его работы, и если оно неудовлетворительное, происходит дополнительное обучение классификатора. Так продолжается до тех пор, пока не будет достигнут требуемый уровень качества или станет ясно, что выбранный алгоритм не работает корректно с данными, либо же данные не имеют структуры, которые можно выявить. К этому типу задач относятся задачи *классификации* и *регрессии.*
   Unsupervised learning выявляет задачи, выявляющие описательные модели. Например закономерности в покупах, совершаемые в больших магазинах. Достоинством таких задач является возможность их решения без каких-либо предварительных знаний об анализируемых данных. К этому типы относится задача *кластеризации.*

2. **Переобучение** – хорошее качество алгоритма на обучении, плохое на новых данных.
   Недообучение – плохое качество на обучающей и на тестовой выборках.
   При этом с недообучением понятно как бороться: нужно усложнять семейство алгоритмов, брать более сложные алгоритмы, например, многочлены высокой степени вместо линейных.
   При проверке на переобучение можно откладывать часть обучающей выборки и на ней проверять наш алгоритм. Существует еще метод кросс-валидации – чуть более сложная проверка отложенной выборки.

3. **Обучающая выборка** – выборка, по которой производится настройка алгоритма(оптимизация параметров).
   Тестовая выборка – выборка, по которой оценивается качество построенной модели.
   Кросс-валидация(cross-validation) – метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на $k-1$ частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования.

В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

0.2 Простые методы

1. kNN кластеризация – задача отнесения объекта к одному из заранее определенных классов на основании его формализованных признаков. Каждый из объектов представляется в виде вектора N-мерного пространства, каждое измерение которого представляет собой описание признаков объекта. Для обучения классификатора необходимо иметь набор объектов, для которых заранее определены классы. Такой набор объектов называют обучающая выборка, её разметка производится вручную.
   Алгоритм
   Для классификации каждого объекта из выборки необходимо проделать следующие операции:
   - Вычислить расстояние до каждого объекта обучающей выборки
   - Отобрать k объектов, расстояние до которых минимально
   - Класс классифицируемого объекта – это класс, наиболее часто встречающий среди k ближайших соседей

0.3 Наивный байес и центроидный классификатор

$P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}$

$a(x) = arg\max_y P(y)\prod_{k=1}^n P(x^{(k)}|y) = arg\max_y[\ln[P(y)\prod_{k=1}^n P(x^{(k)}|y)]] = arg\max_y[\ln[P(y) + \sum_{k=1}^n \ln(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}))]] = arg\max_y[\ln[P(y) + \sum_{k=1}^n \ln[\frac{1}{\sqrt{2\pi\sigma^2}}] + \sum_{k=1}^n(-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2})] = arg\max_y[\ln[P(y) + n\ln[\frac{1}{\sqrt{2\pi\sigma^2}}] + \sum_{k=1}^n(-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2})] = arg\max_y[\ln[P(y) + n\ln[\frac{1}{\sqrt{2\pi\sigma^2}}] - \frac{1}{2\sigma^2}\sum_{k=1}^n(x^{(k)} - \mu_{yk})^2]$

Классы имеют одинаковые априорные вероятности $\Rightarrow P(y) = const$ и $a(x) = arg\min_y \frac{1}{2\sigma^2}\sum_{k=1}^n(x^{(k)} - \mu_{yk})^2 = argmin_y||x - \mu_y||$
Получаем центроидный классификатор

0.4 Ошибка 1NN и оптимального байессовского классификатора

Вероятность ошибки байессовского классификатора: $E_B = minP(1|x); P(0|x) = 1 - max_{y\in 0,1}P(y|x)$

Пусть $r = max_{y \in 0,1} P(y|x)$ Ошибка метода ближайшего соседа $E_{N,n} = P(y \neq y_n)$ При $n \to \infty$ распределение вероятностей классов для ближайшего соседа $x - P(y_n|x_n)_{y_n=0}^{1}$ стремиться к распределению для $x : P(y|x)_{y=0}^{1}$

$E_{N,n} = |$независимость принадлежностей классов для x и $x_n|$ $= \sum_{y \neq y_n \in 0,1} P(y|x)P(y_n|x_n) \to \sum_{y \in 0,1 P(y|x)(1-P(y|x))} = 2r(1-r) \leq 2(1-r) = 2E_B$