Вероятностни и статистически методи за разпознаване. Подход на Бейс. Гаусово разпределение и разстояние по Мехаланобис. Методи за разпознаване по разстояние

В статистическото разпознаване на признаци, един признак се представя чрез множество от *k* характеристики (или атрибути) представени като *k*-мерен вектор. Статистическите подходи се използват, за да се определят граници между класовете на признаците.

# Подход на Бейс

В Бейсовите подходи, ние се стремим да определим най-добрата хипотеза, от известни ни хипотези, при зададени предварително данни и налични данни за априорните вероятности на хипотезите. Алтернативно, при зададени примери и техните класове и априорните вероятности на класовете ние искаме да можем да класифицираме нови примери. Нека имаме входни данни *{xj}* и тяхното разпределение по класовете *{Ci}*, целта ни е да можем да намерим най-правдоподобния клас за даден тестов пример *y*, т.е. да можем да отговаряме на заявката:

Тъй като не можем да отговорим на заявката в явен вид използвайки формулата на Бейс, можем да преобразуваме заявката в по-удобна форма:

Където, *P(Ci)* е априорната вероятност на класа *Ci*. Ако не разполагаме с такава оценка, то можем или да приемем еднаква за всеки клас или да изчислим, базирайки се на броя на елементите в класа. *P(y)* е априорната вероятност да наблюдаваме *y*. Макар че, не можем да я изчислим, тя приема една и съща стойност за всяка хипотеза. Съответно остава единствено да дефинираме метод за изчисляване на *P(y|Ci)*.

Преди да го дефинираме, нека се опитаме да направим оценка на риска. Ако наблюдаваме дадено *y* и го класифицираме с *Ci*и извършим действието *αi*, но то принадлежи на класа *Cy*, каква загуба търпим? По дефиниция загубата е *λ(αi | Cy)*.От гледна точка на теорията на решенията очакваната загуба се нарича *риск*, а *R(αi | y)*се нарича *условен риск*. Формално, задачата се свежда до минимизиране на риска. Можем лесно да изчислим риска използвайки условния риск:

Ще разгледаме тривиалния случай, в който условният риск е един и същ, т.е. най-правдоподобната хипотеза носи най-малък риск.

# Гаусово разпределение

Гаусовото разпределение е непрекъснато вероятностно разпределение, описващо разпределение концентрирано около средното. Графиката на плътността му има формата на камбана, центрирана в средната стойност и е известна като функция на Гаус.

Функцията на плътността на Гаусовото разпределение е:

където *µ* е средната стойност, а *σ* е отклонението.

Функцията на плътността има три много важни свойства, а именно:

* Симетрия относно средната стойност *µ*
* Медианата и модата са равни на *µ*
* Инфлексните точки на кривата са на едно стандартно отклонение от средното

# Разстояние по Мехаланобис

Разстоянието по Мехаланобис се базира на корелацията между променливи, чрез която могат да се идентифицират и анализират връзките между променливата. То се различава от Евклидовото разстояние по това, че то взема предвид корелацията между данните и инвариантно относно мащаба.

Формално, разстоянието по Мехаланобис на*N*-мерния вектор от група от стойности със средни стойности и матрица на ковариацията *S*, се дефинира като:

Естествено, разстоянието може да се дефинира и за два случайни вектора *x*и*y:*

В частния случай, когато матрицата на ковариацията е единичната матрица, разстоянието по Мехаланобис се свежда до Евклидово разстояние. Ако матрицата на ковариацията е диагонална, полученото разстояние се нарича *нормализирано Евклидово разстояние*:

където *si* е стандартното отклонение на *xi*и*yi*.



**Фигура 1: При показаното разпределение, точките по границата на множеството са на еднакво разстояние по Мехаланобис от средното, но на различно Евклидово разстояние.**

# Гаусово разпределение и разстояние по Мехаланобис

Ако предположим, че всеки един от класовете имат нормално разпределение около някакъв център, то:

Тогава:

От това непосредствено следва, че можем да използваме разстояние на Мехаланобис за оценка на *P(x|C)*:

# Методи за разпознаване по разстояние

За зададен пример *y*, зададено обучаващо множество *{xj­}* разпределено в класовете *{Ci}*, искаме да намерим към кой клас*Cy* принадлежи примера.

**Най-близък център (Nearest Centroid)**

За всеки клас*Ci* определяме представител (центроид на класа):

Избираме класа, чийто център е най-близо до примера (с разстояние по Мехаланобис):

Съществената слабост на този подход си проличава при следното разпределение (и Евклидово разстояние), където *y* неправилно ще бъде класифициран като представител на *C2*:



**Най-близък съсед (Nearest Neighbour)**

Класифицираме *y* с класа на най-близкия обучаващ пример:

**kНай-близки съседа (k-Nearest Neighbours)**

Изчислява се разстоянието от *y* до всички *xi*, след което се вземат *k*-те най-близки съседа. Използвайки най-близките съседи можем да определим класа по няколко подхода. Първия е да дадем тегла на класовете на базата на броя на представителите им в най-близките съседи. В редица случаи обаче (например наличието на класове с представители значително по-малко от *k* или някои по-специфични форми на класовете) е по-подходящо да изчислим теглата на класовете като средното разстояние между *y*и представителите на този клас принадлежащи на най-близките съседи. При така изчислените тегла, избираме класа с най-малко тегло.

**k Най-близки центъра (k-Nearest Centers)**

При този подход, за всеки клас изчисляват определен брой центроиди, като целта е те да описват максимално добре формата на класа. Броят на центроидите може да зависи от размера на класа, от формата на класа (прилагане на клъстериращ алгоритъм върху самия клас с цел определяне на оптималния брои центроиди) или да бъде един и същ за всеки клас.

Центроидите могат да бъдат избрани от самите представители на класа чрез итеративно избиране на центроид максимизиращ разстоянието между всички избрани до момента центроиди или чрез клъстериращ алгоритъм.

След избирането на центроидите алгоритъма следва логиката на k-Nearest Neighbours.

# Източници

Scott Ferson; Bayesian methods in risk assessment

Генадий Агре; Лекции по Машинно Самообучение

Wikimedia Foundation, http://en.wikipedia.org