Метод максимального правдоподобия

порождает критерии качества

1 ***Максимальное правдоподобие в случае дискретного распределения.***

Пусть X - случайная величина с дискретным распределением.

Определение

Распределение дискретное, если известны все значения $a_1, a_2, a_3, \ldots, a_k$, которые она может принять и все вероятности $p_1, p_2, p_3, \ldots, p_k$ того, что

$P\{X = a_i\} = p_i$.

Комментарий. Определение для простейшего случая, значений может быть бесконечно много (но счетно), распределение может быть абсолютно непрерывным.

Распределение известно с точностью до нескольких параметров.

Известны независимые наблюдения случайной величины X

$x_1, x_2, x_3, \ldots, x_n$

Определение. Вероятность наблюдать именно те значения, которые были зарегистрированы, называют функцией правдоподобия ***L***.

Вероятность того, что будут наблюдаться именно эти значения.

$$L = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) = .$$

$$. = \prod_{i=1}^{n} P(X_i = x_i)$$

Вопрос 1 : откуда появились n случайных величин?

Вопрос 2 : какие аргументы у функции ***L***?

Вопрос 3 : Использовалась независимость наблюдений. Сохраняется ли независимость после стандартизации?

2 Важный пример. Функция правдоподобия для распределения Бернулли.

Рассматриваем распределение Бернулли,

$$P(X=1)=p$$
$$P(X=0)=1-p$$

Замечание. Неизвестный параметр распределения в этой задаче - p.

$$\textit{правдоподобие}=\mathbf{L}=p^k\cdot(1-p)^{(n-k)}$$

$$\textit{лог правдоподобие}=\log\mathbf{L}=k\cdot\log(p)+(n-k)\cdot\log(1-p)$$

Вопрос. Где используются наблюдения случайных величин?

Выкладки.

$$L=P(X_1=x_1,X_2=x_2,X_3=x_3,\ldots,X_n=x_n)=.$$

$$.=\prod_{i=1}^{n}P(X_i=x_i)=.$$

$$.=p^{x_1}\cdot(1-p)^{(1-x_1)}\cdot p^{x_2}\cdot(1-p)^{(1-x_2)}\cdot p^{x_3}\cdot(1-p)^{(1-x_3)}\cdot\ldots\cdot p^{x_n}\cdot(1-p)^{(1-x_n)}$$

Заметим, что

каждый $x_i, i=1,2,\ldots,n$ равен либо 0, либо 1

каждый $p^{x_i}\cdot(1-p)^{(1-x_i)}$ равно либо $\boldsymbol{p}$, либо 1 - $\boldsymbol{p}$

Привычнее минимизировать критерий качества, а не максимизировать…

Поэтому рассматривают

$$-\textit{лог правдоподобие}=-(k\cdot\log(p)+(n-k)\cdot\log(1-p))$$

3 Logarithmic Loss

Задача классификации, два класса, коды классов 0 и 1.

$y_i, i=1,2,3,\ldots,n$ - значения распознаваемой переменной

$f(x_i), i=1,2,3,\ldots,n$ - выходные значения модели, часто интерпретируются как вероятность принадлежать классу “1”.

$$-\frac{1}{n}\cdot\sum_{i=1}^{n}[y_i\cdot\log(f(x_i))+(1-y_i)\cdot\log(1-f(x_i))]$$

Для наглядного сравнения перепишем

$$-\text{лог правдоподобие}=-(k\cdot\log(p)+(n-k)\cdot\log(1-p))$$

$$.=-(\sum_{i=1}^{n}x_i\cdot\log(p)+\sum_{i=1}^{n}(1-x_i)\cdot\log(1-p))=.$$

$$.=-\sum_{i=1}^{n}[x_i\cdot\log(p)+(1-x_i)\cdot\log(1-p)]$$

О терминологии.

В чем разница между Log Loss и Cross-Entropy?
Разницы нет.

Важные частные случаи см.
Ridgeway
Generalized Boosted Models. A guide to the gbm package
2007 pdf