

逻辑斯蒂回归模型

逻辑回归模型

总概

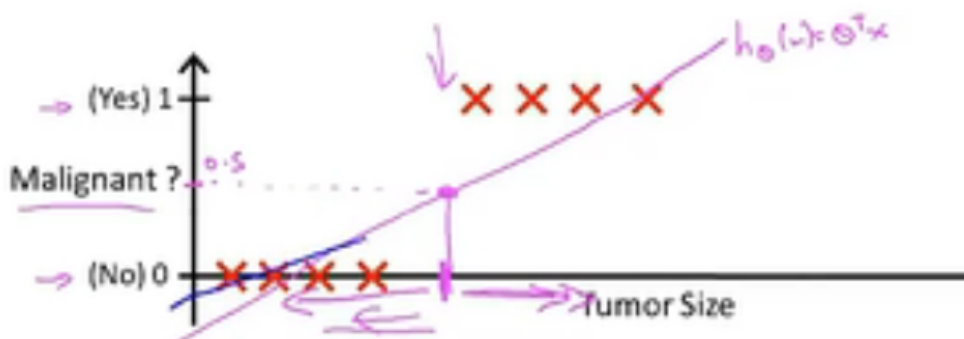
从线性分类开始讲起：逻辑回归是对 $P(y | x)$ 进行建模的模型。它的作用是**分类**。

引入

从线性回归到分类：**为什么不直接进行线性回归呢**

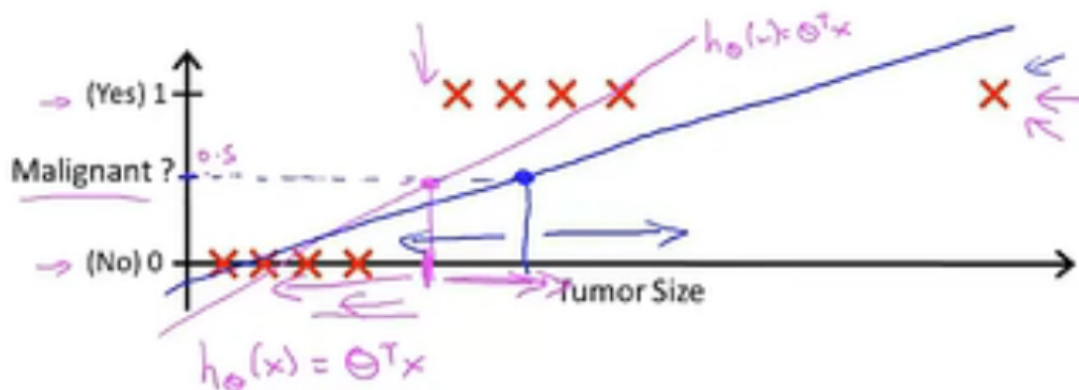
如果已有一个训练好的简单的线性回归的模型，我们可以利用该模型进行预测，假设已知有测试数据集，我们可以拟合出一个图形来预测数据特征值和标签的关系。

对于普通的线性回归，举例：已知肿瘤大小判断其是否为恶性肿瘤。假设现在已经通过线性拟合，拟合出一条表示二者关系的直线。



假设有一个阈值0.5，其对应横坐标为 x_0 ，如果 $x > x_0$ ，那么纵坐标大于0.5超过阈值，判为恶性，反之为良性。目前来看根据粉色线条还可以相对准确的对肿瘤分类。

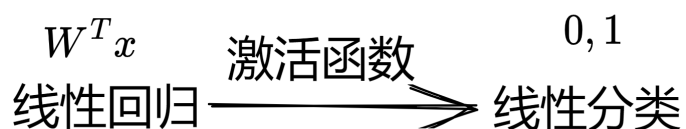
但是，如果有一个和原有的数据集相对较远的数据 x 呢？



如上图，此时根据新的数据集产生了一条新的拟合曲线（如蓝线），如果还按照以往的方式进行分类，效果就很糟。这个新点使得判断的界限从原本相对准确的粉色线挪动到不准确的蓝色线处。

综上，直接将线性回归应用到分类并不是很好的方法。

除此之外，我们还需要一个新的改进模型可以将大于1或者小于0的预测值投射到0到1的区间内，更直观的展现其概率。



从线性回归的组合映射到线性分类的0, 1概率上，使得模型输出一个概率值。逻辑回归就是采用特定的激活函数（sigmoid函数）来完成映射。

怎么找到这个映射？

在总概中提到，逻辑斯蒂回归是针对 $P(y | x)$ 进行建模。

根据条件概率的定义：

$$P(y | x) = \frac{P(y \cap x)}{P(x)}$$

$$P(x | y) = \frac{P(y \cap x)}{P(y)}$$

整理与合并这两个方程式，我们可以得到：

$$P(y | x)P(x) = P(x \cap y) = P(x | y)P(y)$$

两边同时除以 $P(x)$ ，得到贝叶斯公式如下：

$$p(y | x) = \frac{p(x | y)p(y)}{p(x)}$$

$$P(x) = P(x \cap y) + P(x \cap y^C) = P(x | y)P(y) + P(x | y^C)P(y^C)$$

y^C 是 y 的补集。可以通过如下算式计算 $p(y = 1 | x)$ 和 $p(y = 0 | x)$ 。

$$p(y = 1 | x) = \frac{p(x | y = 1)p(y = 1)}{p(x | y = 1)p(y = 1) + p(x | y = 0)p(y = 0)}$$

如果取 $a = \ln \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}$ ，（8）式上下同时除以

$$p(x | y = 1)p(y = 1)$$

,

$$p(y = 1 | x) = \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}$$

$$\frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)} = e^{-a}$$

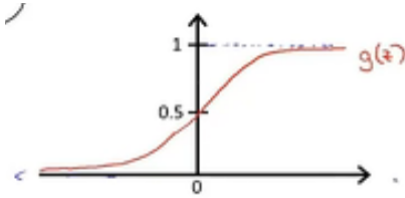
sigmoid函数得到了：

$$p(y = 1 | x) = \frac{1}{1 + e^{-a}}$$

a 就是sigmoid的参数。

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

当 $z \rightarrow \infty$, $\lim \sigma(z) = 1$, 当 $z \rightarrow 0$, $\sigma(z) = \frac{1}{2}$, 当 $z \rightarrow -\infty$, $\lim \sigma(z) = 0$



可以将实数值映射到 $[0,1]$ 区间。

$$R \rightarrow (0,1) \quad w^T x \rightarrow P$$

引入：为什么是条件概率分布？它的含义是什么？

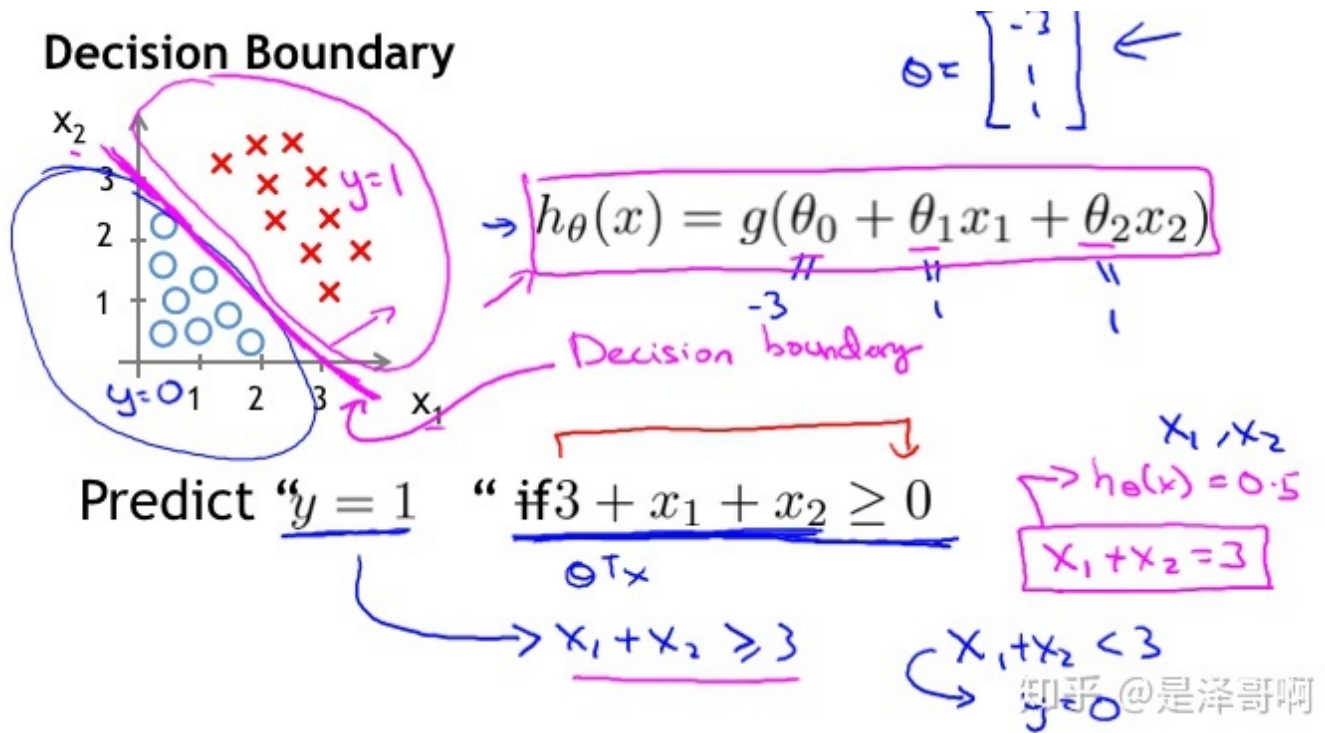
模型输出的是概率 $h_\theta(x)$ ，输出 $y = 1$ 的概率，若仍然以是否为恶性肿瘤举例，如果输出的值为0.7，这说明恶性肿瘤的概率是0.7。

如何表达 $h_\theta(x)$ ？

$$h_\theta(x) = p(y = 1 \mid x; \theta)$$

解释：在给定 x 的情况下，预测 $y=1$ 的概率。病人的特征为 x 的情况下，预测肿瘤为恶性的概率。

Logistic 回归主要用于分类问题，我们以二分类为例，对于所给数据集假设存在这样的一条直线可以将数据完成线性可分。



决策边界可以表示为

$$w_1 x_1 + w_2 x_2 + b = 0$$

，假设某个样本点

$$h_w(x) = w_1 x_1 + w_2 x_2 + b > 0$$

那么可以判断它的类别为 1，这个过程其实是第一章感知机。Logistic 回归还需要加一层，它要找到分类概率 $P(Y=1)$ 与输入向量 x 的直接关系，然后通过比较概率值来判断类别。

$$P_0 : P(y = 1 | x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

$$P_1 : p(y = 0 | x) = 1 - p(y = 1 | x) = 1 - \frac{1}{1 + e^{-w^T x}} = \frac{e^{-w^T x}}{1 + e^{-w^T x}}$$

将两个数学式二合一：

$$P(y | x) = P_1^y P_0^{1-y}$$

补：伯努利分布：

$$f_X(x) = p^x(1-p)^{1-x} = \begin{cases} p & \text{if } x = 1 \\ q & \text{if } x = 0 \end{cases}$$

最大熵原理

最大熵原理是概率模型学习的一个准则。最大熵原理认为，学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型。概率模型的集合一般是约束条件确定的。最大熵原理也就是在所有满足约束条件的模型中选择熵最大的模型。

熵是表示随机变量不确定性的度量，设 X 是一个取有限个值的离散随机变量，其概率分布为

$$P(X = x_i) = p_i, i = 1, 2, \dots, n$$

随机变量 X 的熵定义为：

$$H(P) = - \sum_x P(x) \log P(x)$$

熵满足下列不等式：

$$0 \leq H(P) \leq \log |X|$$

$|X|$ 是 X 的取值个数，当且仅当 X 是均匀分布的时候右边的等号成立。 X 服从均匀分布的时候，熵最大。

举例：

例 6.1 假设随机变量 X 有 5 个取值 $\{A, B, C, D, E\}$, 要估计取各个值的概率 $P(A), P(B), P(C), P(D), P(E)$ 。

解 这些概率值满足以下约束条件:

$$P(A) + P(B) + P(C) + P(D) + P(E) = 1$$

满足这个约束条件的概率分布有无穷多个。如果没有任何其他信息, 仍要对概率分布进行估计, 一个办法就是认为这个分布中取各个值的概率是相等的:

$$P(A) = P(B) = P(C) = P(D) = P(E) = \frac{1}{5}$$

等概率表示了对事实的无知。因为没有更多的信息, 这种判断是合理的。

有时, 能从一些先验知识中得到一些对概率值的约束条件, 例如:

$$P(A) + P(B) = \frac{3}{10}$$

$$P(A) + P(B) + P(C) + P(D) + P(E) = 1$$

满足这两个约束条件的概率分布仍然有无穷多个。在缺少其他信息的情况下, 可以认为 A 与 B 是等概率的, C, D 与 E 是等概率的, 于是,

$$P(A) = P(B) = \frac{3}{20}$$

$$P(C) = P(D) = P(E) = \frac{7}{30}$$

如果还有第 3 个约束条件:

$$P(A) + P(C) = \frac{1}{2}$$

$$P(A) + P(B) = \frac{3}{10}$$

$$P(A) + P(B) + P(C) + P(D) + P(E) = 1$$

可以继续按照满足约束条件下求等概率的方法估计概率分布。这里不再继续讨论。以上概率模型学习的方法正是遵循了最大熵原理。 ■

最大熵模型

假设分类模型是条件分布模型 $P(y | x)$ ， x 为输入， y 为输出。 X 为输入的集合， Y 为输出的集合。这个模型就是输入 x ，以条件概率的规则输出 y 。现有一个数据集 T ，想用最大熵原理选择一个最好的分类模型。

首先应该先确定约束条件：在训练数据集已知的情况下，联合分布 $P(X,Y)$ 和边缘分布 $P(X)$ 是可以确定的。

$$\tilde{P}(X = x, Y = y) = \frac{\nu(X = x, Y = y)}{N}$$

$$\tilde{P}(X = x) = \frac{\nu(X = x)}{N}$$

分子 $\tilde{P}(X = x, Y = y)$ 是样本 (x, y) 出现频数， N 是样本容量。

定义特征函数 $f(x, y)$ 描述输入 x 和输出 y 的某一个事实。定义为：

$$f(x, y) = \begin{cases} 1, & x \text{ 与 } y \text{ 满足某一事实} \\ 0, & \text{否则} \end{cases}$$

期望的算式：

$$E(X) = \sum_i p_i x_i$$

特征函数关于分布 $\tilde{P}(X, Y)$ 的期望值，用下式表示：

$$E_{\tilde{P}}(f) = \sum_{x,y} \tilde{P}(x, y) f(x, y)$$

对于 $P(Y, X)$ 的期望：

$$E_P(f) = \sum_{x,y} P(x, y) f(x, y)$$

$$P(y | x) = \frac{P(x, y)}{P(x)}$$

鉴于我们并不知道 $P(x)$ 的具体分布，所以用 $\tilde{P}(x)$ 来模拟 $P(x)$ 。

$$E_P(f) = \sum_{x,y} \tilde{P}(x)P(y|x)f(x,y)$$

可以假设两个期望值相等（理论的和经验的）：

$$E_P(f) = E_{\tilde{P}}$$

$$\sum_{x,y} \tilde{P}(x)P(y|x)f(x,y) = \sum_{x,y} P(x,y)f(x,y)$$

将 (27) 或者 (28) 作为模型学习的约束条件，假设有 n 个特征函数 $f_i(x,y)$ ，就有 n 个约束条件。

最大熵模型

定义 6.3（最大熵模型） 假设满足所有约束条件的模型集合为

$$\mathcal{C} \equiv \{P \in \mathcal{P} | E_P(f_i) = E_{\tilde{P}}(f_i), \quad i = 1, 2, \dots, n\} \quad (6.12)$$

定义在条件概率分布 $P(Y|X)$ 上的条件熵为

$$H(P) = - \sum_{x,y} \tilde{P}(x)P(y|x) \log P(y|x) \quad (6.13)$$

则模型集合 \mathcal{C} 中条件熵 $H(P)$ 最大的模型称为最大熵模型。式中的对数为自然对数。

条件熵：

条件熵 $H(Y|X)$ 表示在已知随机变量 X 的条件下随机变量 Y 的不确定性。条件熵 $H(Y|X)$ 定义为 X 给定条件下 Y 的条件概率分布的熵对 X 的数学期望：

$$\begin{aligned}
H(Y | X) &= \sum_x p(x) H(Y | X = x) \\
&= - \sum_x p(x) \sum_y p(y | x) \log p(y | x) \\
&= - \sum_x p(x, y) \log p(y | x) \\
&= - \sum_{x,y} p(x, y) \log p(y | x)
\end{aligned}$$

条件概率分布 $P(Y | X)$ 条件熵推导：

$$H(Y | X) = - \sum_{i=1} P(x_i) H(y | x = x_i) = - \sum_{x,y} \tilde{P}(x) P(y | x) \log P(y | x)$$

最大熵模型的学习

最大熵模型的学习就是求解最大熵模型。 **目的：熵最大**

最大熵模型的求解思路和步骤如下：

1. 利用 Lagrange 乘子法将最大熵模型由一个带约束的最优化问题转化为一个与之等价的无约束的最优化问题，它是一个**极小极大问题** (min max).
2. 利用对偶问题等价性，转化为求解上一步得到的极小极大问题的对偶问题，它是一个**极大极小问题** (max min).

在求解内层的极小问题时，可以导出最大熵模型的解具有**指数形式**，而在求解外层的极大问题时，还将意外地发现其与**最大似然估计**的等价性。

https://blog.csdn.net/hgnuxc_1993

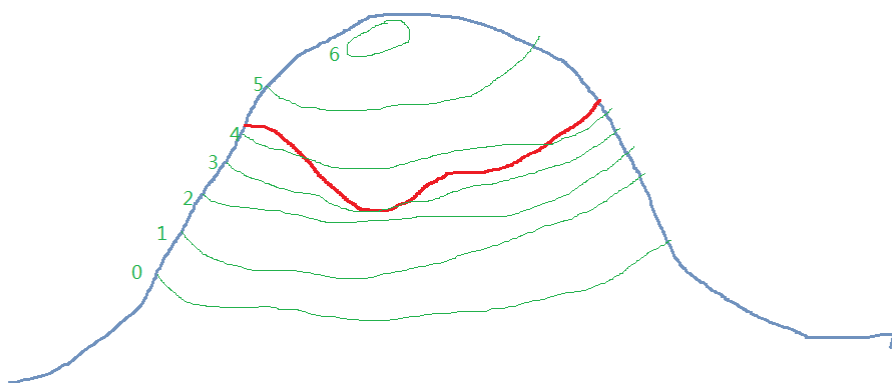
$$\begin{array}{ll}
\max_{P \in \mathbf{C}} & H(P) = - \sum_{x,y} \tilde{P}(x) P(y | x) \log P(y | x) \\
\text{s.t.} & E_P(f_i) = E_{\tilde{P}}(f_i), \quad i=1,2, \dots, n \\
& \sum_y P(y | x) = 1
\end{array}$$
 最大值问题改写为等价最小值问题：

$$\begin{array}{ll}
\min_{P \in \mathbf{C}} & -H(P) = \sum_{x,y} \tilde{P}(x) P(y | x) \log P(y | x) \\
\text{s.t.} & E_P(f_i) - E_{\tilde{P}}(f_i) = 0, \quad i=1,2, \dots, n
\end{array}$$

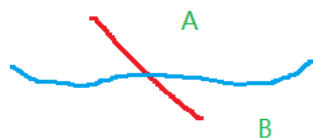
$\sum_y P(y \mid x) = 1$ 求解 (32) : (采用拉格朗日乘子)

拉格朗日乘子的解释:

想象一下, 目标函数 $f(x, y)$ 是一座山的高度, 约束 $g(x, y) = C$ 是镶嵌在山上的一条红线如下图。为了找到曲线上的最低点, 就从最低的等高线 (0那条) 开始往上数。数到第三条, 等高线终于和曲线有交点了 (如上图所示)。因为比这条等高线低的地方都不在约束范围内, 所以这肯定是这条约束曲线的最低点了。



而且约束曲线在这里不可能和等高线相交, 一定是相切。因为如果是相交的话, 如下图所示, 那么曲线一定会有一部分在B区域, 但是B区域比等高线低, 这是不可能的。



两条曲线相切, 意味着他们在这点的法线平行, 也就是法向量只差一个任意的常数乘子。

$$\nabla f(x, y) = -\lambda \nabla g(x, y)$$

$$\nabla(f(x, y) + \lambda g(x, y)) = 0。$$

上述算式就是

$$f(x, y) + \lambda g(x, y)$$

无约束情况下极值点的必要条件。改写一下 (34)：求函数 $f(x, y)$ 在条件

$$\begin{cases} \frac{\partial L}{\partial \mathbf{x}} = \mathbf{0}, \\ g_i(\mathbf{x}) = 0, \quad i=1, 2, \dots, m \end{cases}$$

其中 $\frac{\partial L}{\partial \mathbf{x}} = \left(\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \dots, \frac{\partial L}{\partial x_n} \right)^T$ 表示 L 关于 \mathbf{x} 的梯度。 $f(x) = -$

$$\begin{aligned} L(P, w) &\equiv -H(P) + w \left(1 - \sum_y P(y \mid x) \right) + \sum_{i=1}^n w_i \left(E\{\tilde{P}\} \left(f_i \right) - E_P \left(f_i \right) \right) \\ &= - \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P(y \mid \mathbf{x}) \log P(y \mid \mathbf{x}) + w \left(1 - \sum_y P(y \mid \mathbf{x}) \right) + \\ &\quad + \sum_{i=1}^n w_i \left(\sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) f_i(\mathbf{x}, y) - \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P(y \mid \mathbf{x}) f_i(\mathbf{x}, y) \right) \end{aligned}$$

最优化的原始问题：

$$\min_{P \in \mathbf{C}} \max_w L(P, w)$$

转化为对偶问题：

$$\max_w \min_{P \in \mathbf{C}} L(P, w)$$

如何求解此对偶问题？先求解内层 $\min_{P \in \mathbf{C}} L(P, w)$ ， $\min_{P \in \mathbf{C}} L(P, w)$ 是关于

$$\Psi(w) = \min_{P \in \mathbf{C}} L(P, w) = L(P_w, w)$$

该函数的解为： $P_w = \arg \min_{P \in \mathbf{C}} L(P, w) = P_w(y | x)$ 。如何求解 w 呢？

$$L(p, w) = \sum_{\{x, y\}} \tilde{P}(x) P(y | x) \log P(y | x) + w_0 \left(1 - \sum_y P(y | x) \right) + \sum_{i=1}^n w_i \left(\sum_{\{x, y\}} \tilde{P}(x, y) f_i(x, y) - \sum_{\{x, y\}} \tilde{P}(x) P(y | x) f_i(x, y) \right)$$

$$\begin{aligned} \frac{\partial L(P, w)}{\partial P(y | x)} &= \sum_{\{x, y\}} \tilde{P}(x) (\log P(y | x) + 1) - \sum_y w_0 - \sum_{\{x, y\}} \tilde{P}(x) \sum_{i=1}^n w_i f_i(x, y) \\ &= \sum_{\{x, y\}} \tilde{P}(x) (\log P(y | x) + 1 - w_0 - \sum_{i=1}^n w_i f_i(x, y)) \end{aligned}$$

令偏导数为0，求解 $P(y | x)$ ：

$$P(y | x) = \frac{\exp \left(\sum_{i=1}^n w_i f_i(x, y) + w_0 - 1 \right)}{\exp \left(\sum_{i=1}^n w_i f_i(x, y) \right) \exp \left(1 - w_0 \right)} \\ p(y | x) = e^{w_0 - 1} \cdot e^{\sum_{i=1}^n w_i f_i(x, y)},$$

将上式代入约束条件 $\sum_y p(y | x) = 1$ ，即

$$\sum_y p(y | x) = e^{w_0 - 1} \cdot \sum_y e^{\sum_{i=1}^n w_i f_i(x, y)} = 1,$$

可得

$$e^{w_0-1} = \frac{1}{\sum_y e^{\sum_{i=1}^n w_i f_i(x, y)}} .$$

将(44)代回(42), 得到

$$p_w = \frac{1}{Z_w(x)} e^{\sum_{i=1}^n w_i f_i(x, y)},$$

其中

$$Z_w(x) = \sum_y e^{\sum_{i=1}^n w_i f_i(x, y)}$$

$$\text{当 } Z_w(x) = \sum_y \exp \left(\sum_{i=1}^n w_i f_i(x, y) \right),$$

$$P(y \mid x) = \frac{1}{Z_w(x)} \exp \left(\sum_{i=1}^n w_i f_i(x, y) \right)$$

$Z_w(x)$ 是规范化因子, $f_i(x, y)$ 是特征函数, w_i 是特征函数的权重。

$$\max_w \Psi(w)$$

将45中的解记为 w^* :

$$w^* = \arg \max_w \Psi(w)$$

对于 w^* , 有一个对应的, 需要求解的 P_* , 也即最大熵模型的求解就

$$L(\tilde{P}) - \left(P_w \right) = \log \prod_{x, y} P(y \mid x)^{\tilde{P}(x, y)} = \sum_{x, y} \tilde{P}(x, y) \log P(y \mid x)$$

之前提出的最大熵模型如下:

$$P(y \mid x) = \frac{1}{Z_w(x)} \exp \left(\sum_{i=1}^n w_i f_i(x, y) \right)$$

将 (47) 代入 (48) :

```

\begin{aligned}
L(\tilde{P}) &= \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) \log P(y \mid \mathbf{x}) \\
&= \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) \sum_{i=1}^n w_i f_i(\mathbf{x}, y) - \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) \log Z_w(\mathbf{x}) \\
&= \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) \sum_{i=1}^n w_i f_i(\mathbf{x}, y) - \sum_{\mathbf{x}} \tilde{P}(\mathbf{x}) \log Z_w(\mathbf{x})
\end{aligned}

```

对于对偶函数 $\Psi(w) = \min_{P \in \mathcal{C}} L(P, w) = L(P_w, w)$,

```

L(p, w) = \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P(y \mid \mathbf{x}) \log P(y \mid \mathbf{x}) + w_0 \left( 1 - \sum_y P(y \mid \mathbf{x}) \right) +
\sum_{i=1}^n w_i \left( \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) f_i(\mathbf{x}, y) - \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P(y \mid \mathbf{x}) f_i(\mathbf{x}, y) \right)

```

```

\sum_y P(y \mid \mathbf{x}) = 1

```

```

\begin{aligned}
\Psi(w) &= \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P_w(y \mid \mathbf{x}) \log P_w(y \mid \mathbf{x}) + \\
&\quad \sum_{i=1}^n w_i \left( \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) f_i(\mathbf{x}, y) - \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P_w(y \mid \mathbf{x}) f_i(\mathbf{x}, y) \right) \\
&= \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) \sum_{i=1}^n w_i f_i(\mathbf{x}, y) + \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P_w(y \mid \mathbf{x}) \left( \log P_w(y \mid \mathbf{x}) - \sum_{i=1}^n w_i f_i(\mathbf{x}, y) \right) \\
&= \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}, y) \sum_{i=1}^n w_i f_i(\mathbf{x}, y) - \sum_{\mathbf{x}, y} \tilde{P}(\mathbf{x}) P_w(y \mid \mathbf{x}) \log Z_w(\mathbf{x})
\end{aligned}

```

$$\begin{aligned} & \& = \sum_{\{x, y\}} \tilde{P}(x, y) \sum_{i=1}^n w_i f_i(x, y) - \sum_x \tilde{P}(x) \log Z_w(x) \\ & \end{aligned}$$

比较 (51) 和 (48) :

$$\Psi(w) = L_{\tilde{P}}(P_w)$$

对偶函数的最大化也就等价于对数似然函数的最大化。*要求解的对偶函数

$$P_w(y \mid x) = \frac{1}{Z_w(x)} \exp \left(\sum_{i=1}^n w_i f_i(x, y) \right)$$

$$Z_w(x) = \sum_y \exp \left(\sum_{i=1}^n w_i f_i(x, y) \right)$$

对数似然函数:

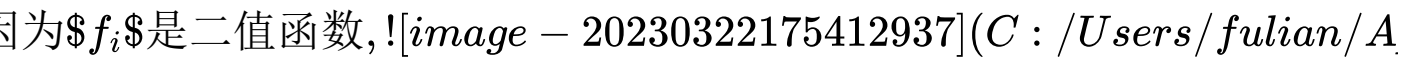
$$\begin{aligned} & \begin{aligned} & L_{\tilde{P}}(P_w) \& = \sum_{\{x, y\}} \tilde{P}(x, y) \\ & \sum_{i=1}^n w_i f_i(x, y) - \sum_x \tilde{P}(x) \log Z_w(x) \end{aligned} \\ & \end{aligned}$$

目标: 找到使得 $L_{\tilde{P}}(P_w)$ 最大的 w , *IIS* 的想法是: 假设最大熵模型当前

$$\begin{aligned} & \begin{aligned} & L(w + \delta) - L(w) \& = \sum_{\{x, y\}} \tilde{P}(x, y) \log P_{w+\delta}(y \\ & \mid x) - \sum_{\{x, y\}} \tilde{P}(x, y) \log P_w(y \mid x) \& \\ & = \sum_{\{x, y\}} \tilde{P}(x, y) \sum_{i=1}^n \delta_i f_i(x, y) - \\ & \sum_x \tilde{P}(x) \log \frac{Z_{w+\delta}(x)}{Z_w(x)} \end{aligned} \\ & \end{aligned}$$

< imgsrc = "C : /Users/fulian/AppData/Roaming/Typora/typora - use

$$f^{\{ \# \}}(x, y) = \sum_i f_i(x, y)$$

因为 f_i 是二值函数,  (C : /Users/fulian/A

$$\begin{aligned} A(\delta \mid w) &= \sum_{\{x, y\} \in \tilde{P}} \sum_{i=1}^n \delta_{f_i(x, y)+1-} \\ &\quad \sum_x \tilde{P}(x) \sum_y P_w(y \mid x) \exp \left(f^{\{ \# \}}(x, y) \sum_{i=1}^n \frac{\delta_{f_i(x, y)}}{f^{\{ \# \}}(x, y)} \right) \end{aligned}$$

对任意 i , 有 $\frac{f_i(x, y)}{f^{\#}(x, y)} \geq 0$ 且 $\sum_{i=1}^n \frac{f_i(x, y)}{f^{\#}(x, y)} = 1$ 这一事实, 根据 *Jensen*

$$\exp \left(\sum_{i=1}^n \frac{f_i(x, y)}{f^{\#}(x, y)} \delta_{f_i(x, y)} \right) \leq \sum_{i=1}^n \frac{f_i(x, y)}{f^{\#}(x, y)} \exp \left(\delta_{f_i(x, y)} \right)$$

注: *Jensen*不等式是:

$$\varphi \left(\sum_{i=1}^n g(x_i) \lambda_i \right) \leq \sum_{i=1}^n \varphi(g(x_i)) \lambda_i$$

于是式(58)可改写为

$$\begin{aligned} A(\delta \mid w) &\geq \sum_{\{x, y\} \in \tilde{P}} \sum_{i=1}^n \delta_{f_i(x, y)+1-} \\ &\quad \sum_x \tilde{P}(x) \sum_y P_w(y \mid x) \sum_{i=1}^n \left(\frac{f_i(x, y)}{f^{\#}(x, y)} \right) \exp \left(\delta_{f_i(x, y)} \right) \end{aligned}$$

记不等式(61)右端为

$$B(\delta | w) = \sum_{\{x, y\}} \tilde{P}(x, y) \sum_{i=1}^n \delta_i f_i(x, y) - \sum_x \tilde{P}(x) \sum_y P_w(y | x) \sum_{i=1}^n \left(\frac{f_i(x, y)}{f^{\#}(x, y)} \right) \exp \left(\delta_i f^{\#}(x, y) \right)$$

于是得到

$$L(w + \delta) - L(w) \geq B(\delta | w)$$

这里, $B(\delta | w)$ 是对数似然函数改变量的一个新的(相对不紧的)下界。求

$$\frac{\partial B(\delta | w)}{\partial \delta_i} = \sum_{\{x, y\}} \tilde{P}(x, y) f_i(x, y) - \sum_x \tilde{P}(x) \sum_y P_w(y | x) f_i(x, y) \exp \left(\delta_i f^{\#}(x, y) \right)$$

在式(64)里, 除 δ_i 外不含任何其他变量。令偏导数为0得到 δ_i , 依次对 δ_i ,

$$\frac{\partial f(w)}{\partial w_i} = \sum_{\{x, y\}} \tilde{P}(x) P_w(y | x) f_i(x, y) - E\{\tilde{P}\} \left(f_i \right), \quad i=1, 2, \dots, n$$