算法7.1（最大间隔法）

输入：线性可分训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，
其中 $x_i\in\mathcal{X}=R^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$；

输出：最大间隔分离超平面和分类决策函数。

(1) 构造并求解约束最优化问题：

$$\min_{w,b}\ \frac{1}{2}\|w\|^2$$

$$s.t.\quad y_i(w\cdot x_i+b)-1\geq 0,\ i=1,2,\cdots,N.$$

求得最优解 w^*,b^*。

(2) 由此得到分离超平面：

$$w^*\cdot x+b^*=0.$$

分类决策函数 $f(x)=\mathrm{sign}(w^*\cdot x+b^*)$.

算法7.2（线性可分支持向量机学习算法）

(1) 构造并求解约束最优化问题：

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\alpha_i\geq 0,\ i=1,2,\cdots,N.$$

求得最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^T$.

(2) 计算：

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$$

并选择 α^* 的一个正分量 $\alpha_j^*>0$，计算

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j).$$

(3) 得到分离超平面：

$$w^*\cdot x+b^*=0.$$

分类决策函数 $f(x)=\mathrm{sign}(w^*\cdot x+b^*)$.

算法7.3（软间隔最大化）

输出：分离超平面和分类决策函数

(1) 选择惩罚参数 $C>0$.

构造并求解凸二次规划问题

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0.$$

$$0\leq\alpha_i\leq C,\ i=1,2,\cdots,N.$$

求得最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^T$.

(2) 计算

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i.$$

选择 α^* 的一个分量 α_j^* 适合条件 $0<\alpha_j^*<C$，计算：

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j)$$

(3). ……

算法7.4（非线性支持向量机学习算法）

输出：分类决策函数。

(1) 选取适当的核函数 $K(x,z)$ 和适当的参数 C，构造最优化问题：

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0.$$

$$0\leq\alpha_i\leq C,\ i=1,2,\cdots,N.$$

求得最优解：$\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^T$.

(2) 选择 α^* 的一个正分量 $0<\alpha_j^*<C$，计算

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i K(x_i\cdot x_j).$$

(3) 构造决策函数：

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i K(x\cdot x_i)+b^*\right).$$

1. 只有第一个算法的最优化问题是"$\min_{w,b}\frac{1}{2}\|w\|^2$" $s.t.\ y_i(w\cdot x_i+b)-1\geq 0$.
2. 只有最后一个算法是 $f(x)=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i K(x\cdot x_i)+b^*\right)$
3. 第一个求 w^*,b^*，其余求 α^*，第三、四个有 $\leq C$ 的约束。
4. 2和3区别只有C，剩下的都一样，这也是老师为什么说"三个算法……

感知机学习算法的原始形式：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in\mathcal{X}=R^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$；学习率 $\eta\ (0<\eta\leq 1)$；

输出：w,b；感知机模型 $f(x)=\mathrm{sign}(w\cdot x+b)$.

(1) 选取初值 w_0,b_0。

(2) 在训练集中选取数据 (x_i,y_i).

(3) 如果 $y_i(w\cdot x_i+b)\leq 0$，

$$w\leftarrow w+\eta y_i x_i$$

$$b\leftarrow b+\eta y_i$$

(4) 转至(2)，直至训练集中没有误分类点。

对偶形式：

输出：α,b；感知机模型 $f(x)=\mathrm{sign}\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x+b\right)$.

其中 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)^T$.

(1) $\alpha\leftarrow 0$，$b\leftarrow 0$.

(2) 在训练集中选取数据 (x_i,y_i)

(3) 如果 $y_i\left(\sum_{j=1}^{N}\alpha_j y_j x_j\cdot x_i+b\right)\leq 0$.

$$\alpha_i\leftarrow\alpha_i+\eta$$

$$b\leftarrow b+\eta y_i$$

(4) 转至(2)直到没有误分类数据。

感知机模型：$f(x)=\mathrm{sign}(w\cdot x+b)$.

学习策略：极小化损失函数 —— $\min_{w,b}L(w,b)=-\sum_{x_i\in M}y_i(w\cdot x_i+b)$.

梯度下降法不断极小化目标函数。

第四章 朴素贝叶斯法

4.1.1 基本方法：

先验概率分布 $P(Y=c_k)$，$k=1,2,\cdots,K$.

条件概率分布 $P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$.

参数个数为 $K\prod_{j=1}^{n}S_j$.

条件独立性假设：$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$

$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k).$$

朴素贝叶斯会牺牲一定的分类准确率

后验概率分布：$P(Y=c_k|X=x)=\dfrac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)}$

朴素贝叶斯分类器：$y=f(x)=\arg\max_{c_k}\dfrac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}$

$$y=f(x)=\arg\max_{c_k}P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)$$

4.1.2 后验概率最大化的含义：

选择0-1损失函数：$L(Y,f(X))=\begin{cases}1, & Y\neq f(X)\\ 0, & Y=f(X)\end{cases}$.

为了使期望最小化，只需对 $X=x$ 逐个极小化，由此得到：

$$f(x)=\arg\min_{y\in\mathcal{Y}}\sum_{k=1}^{K}L(c_k,y)P(c_k|X=x)$$

$$=\arg\min_{y\in\mathcal{Y}}\sum_{k=1}^{K}(1-P(y=c_k|X=x)).$$

$$=\arg\max_{y\in\mathcal{Y}}P(y=c_k|X=x).$$

4.2.1. 极大似然估计：

先验概率极大似然估计 $P(Y=c_k)=\dfrac{\sum_{i=1}^{N}I(y_i=c_k)}{N}$

条件概率：$P(X^{(j)}=a_{jl}|Y=c_k)=\dfrac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)}$.

算法4.1

输入：训练数据 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^T$，$x_i^{(j)}$ 是第 j 个样本的第 j 个特征，$x_i^{(j)}\in\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$，$a_{jl}$ 是第 j 个特征可能取的第 l 个值，$i=1,2,\cdots,N$，$j=1,2,\cdots,n$，$l=1,2,\cdots,S_j$，$y_i\in\{c_1,c_2,\cdots,c_K\}$；

输出：实例 x 的分类。

(1) 计算先验概率及条件概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N},\quad k=1,2,\cdots,K.$$

$$P(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)}$$

$j=1,2,\cdots,n$；$l=1,2,\cdots,S_j$；$k=1,2,\cdots,K$.

(2) 对于给定的实例 $x=(x^{(1)},x^{(2)},\cdots,x^{(n)})^T$，计算

$$P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\mid Y=c_k),$$

(3) 确定实例 x 的类

$$y=f(x)=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\mid Y=c_k).$$

学习朴素贝叶斯分类器并确定类标记 y.

4.2.3 贝叶斯估计：

条件概率：$P_\lambda(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda}$.

$\lambda=1$ → 拉普拉斯平滑.

$P_\lambda(X^{(j)}=a_{jl}\mid Y=c_k)>0$. $\sum_{l=1}^{S_j}P(X^{(j)}=a_{jl}\mid Y=c_k)=1$

最大间隔分离超平面的存在唯一性

若训练数据集 T 线性可分，则可将训练数据集中的样本点完全正确地分开的最大间隔分离超平面存在且唯一.

证明：(1) 存在性

∵训练数据集线性可分

∴最优化问题 $\min_{w,b}\ \frac{1}{2}\|w\|^2$ —— 凸函数的极小值解

$s.t.\ y_i(w\cdot x+b)-1\geq 0.$

又∵目标函数有下界.

∴上述最优化问题必有解，记作 (w^*,b^*).

∵训练数据集中又有正实点，又有负实点.

∴$(w,b)=(0,b)$ 不是可行解

∴(w^*,b^*) 必满足 $w^*\neq 0$. 故分离超平面可行.

(2) 唯一性

① 首先证明 w^* 的唯一性. 假设有两个最优解 $(w^*,b_1^*),(w^*,b_2^*)$.

更新 $\|w^*\|=\|w_3\|=C$. 令 $w=\frac{w_2^*+w_3^*}{2}$，$b=\frac{b_1^*+b_2^*}{2}$.

得知 (w,b) 是可行解，从而有：

$$C\leq\|w\|\leq\frac{1}{2}\|w_2^*\|+\frac{1}{2}\|w_3^*\|=C.$$

由等号可得：$\|w\|=\frac{1}{2}\|w_2^*\|+\frac{1}{2}\|w_3^*\|$，从而 $w_2^*=\lambda w_3^*$.

$|\lambda|=1$. 若 $\lambda=-1$，$w=0$，(w,b) 不是解，矛盾. 则 $\lambda=1$，$w_2^*=w_3^*$.

② 设 x_1' 和 x_2' 是集合 $\{x_i\mid y_i=+1\}$ 中分别对应于 $(w^*,b_1^*),(w^*,b_2^*)$ 使 $y_i(w\cdot x_i+b)=1$ 成立的点，x_1'' 和 x_2'' 是集合 $\{x_i\mid y_i=-1\}$ 中分别对应于 (w^*,b_1^*) 和 (w^*,b_2^*) 使 $\cdots$ 成立的点.

得：$b_1^*-b_2^*=-\frac{1}{2}[(w^*\cdot x_1'-x_2')+(w^*\cdot x_1''-x_2'')]$.

又 $w^*\cdot x_2'+b_1^*\geq 1=w^*\cdot x_1'+b_1^*$.

$w^*\cdot x_1'+b_2^*\geq 1=w^*\cdot x_2'+b_2^*$

∴$w^*\cdot(x_1'-x_2')=0$. 同理 $w^*\cdot(x_1''-x_2'')=0$. $\Rightarrow b_1^*=b_2^*$. 得证.

拉格朗日函数：

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i.$$

$$\min_{w,b}L(w,b,\alpha)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(x_i\cdot x_j)y_iy_j+\sum_{i=1}^{N}\alpha_i.$$

$$\max_{\alpha}-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i.$$

对偶问题

等价于极小问题：$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i.$

$$\begin{cases}w^*=\sum_{i=1}^{N}\alpha_i^*y_i\cdot x_i\\ b^*=y_i-\sum_{i=1}^{N}\alpha_i^*y_i(x_i\cdot x_j)\end{cases}$$ —— 原始问题的解

分离超平面可以写成 $\sum_{i=1}^{N}\alpha_iy_i(x\cdot x_i)+b^*=0$.

分类决策函数 $f(x)=sign(\sum_{i=1}^{N}\alpha_i^*y_i(x\cdot x_i)+b^*)$.

核函数：$\mathcal{X}$ 是输入空间，$\mathcal{H}$ 是特征空间，存在映射 $\phi(x):\mathcal{X}\to\mathcal{H}$

对取 $K(x,z)=\phi(x)\cdot\phi(z)$.

那么称 $K(x,z)$ 为核函数，$\phi(x)$ 为映射函数.

$W(\alpha)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_jK(x_i,x_j)-\sum_{i=1}^{N}\alpha_i.$

分类决策函数：$f(x)=sign(\sum_{i=1}^{N}\alpha_i^*y_iK(x_i,x)+b^*)$

常用核函数：

多项式核函数：$K(x,z)=(x\cdot z+1)^p$.

高斯核函数：$K(x,z)=\exp\left(\frac{\|x-z\|^2}{2\sigma^2}\right)$.

7.4. $g(x)=\sum_{i=1}^{N}\alpha_iy_iK(x_i,x)+b.$

$E_i=g(x_i)-y_i=(\sum_{j=1}^{N}\alpha_jy_jK(x_i,x)+b)-y_i.$

SMO：选违反KKT条件最严重的样本点.

0为有足够大的变化.

7.5. SMO算法.

输入：精度 ε；

输出：近似解 $\hat{\alpha}$.

(1) 取初值 $\alpha^{(0)}=0$，令 $k=0$；

(2) 选取优化变量 $\alpha_1^{(k)},\alpha_2^{(k)}$，解析求解最优化问题，得到 $\alpha_1^{(k+1)},\alpha_2^{(k+1)}$，更新 α 为 $\alpha^{(k+1)}$.

(3) 若在精度 ε 范围内满足停机条件

$$\sum_{i=1}^{N}\alpha_iy_i=0.$$

$$0\leq\alpha_i\leq C.\quad i=1,2,\cdots,N.$$

$$y_ig(x_i)=\begin{cases}\geq 1 & \text{若 } \alpha_i=0\\ =1 & \text{若 } 0<\alpha_i<C\\ \leq 1 & \text{若 } \alpha_i=C\end{cases}$$

其中，$g(x_i)=\sum_{j=1}^{N}\alpha_jy_jK(x_j,x_i)+b.$

则转(4)；否则令 $k=k+1$，转(2)；

(4) 取 $\hat{\alpha}=\alpha^{(k+1)}$.

补：最小化合页损失函数的合页损失函数：$\sum_{i=1}^{N}[1-y_i(w\cdot x_i+b)]_++\lambda\|w\|^2$.

拉格朗日对偶问题：$L(w,b,\xi,\alpha,\mu)$

$$=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\alpha_i[y_i(w\cdot x_i+b)+\xi_i-1]-\sum_{i=1}^{N}\mu_i\xi_i$$

原始：min max. 对偶 max min.

6.1 逻辑斯蒂分布

分布函数 $F(x)=P(X\le x)=\frac{1}{1+e^{-(x-\mu)/\gamma}}$

二项逻辑斯蒂回归模型：

$$\begin{cases} P(Y=1|x)=\dfrac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)} \\ P(Y=0|x)=\dfrac{1}{1+\exp(w\cdot x+b)} \end{cases}$$

将权值向量和输入向量扩充

$$\begin{cases} P(Y=1|x)=\dfrac{\exp(w\cdot x)}{1+\exp(w\cdot x)} \\ P(Y=0|x)=\dfrac{1}{1+\exp(w\cdot x)} \end{cases}$$

几率：事件发生与事件不发生概率之比：$\frac{p}{1-p}$ 对数几率：$\log\frac{p}{1-p}$

对数模型：$\log\frac{P(Y=1|x)}{1-P(Y=1|x)}=w\cdot x$

模型的参数估计：(应用极大似然估计估计参数)

设：$P(Y=1|x)=\pi(x)$，$P(Y=0|x)=1-\pi(x)$

似然函数：$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$

对数似然函数：$L(w)=\sum_{i=1}^{N}[y_i\log\pi(x_i)+(1-y_i)\log(1-\pi(x_i))]$

$$=\sum_{i=1}^{N}\left[y_i\log\frac{\pi(x_i)}{1-\pi(x_i)}+\log(1-\pi(x_i))\right]$$

$$=\sum_{i=1}^{N}[y_i(w\cdot x_i)-\log(1+\exp(w\cdot x_i))]$$

对 $L(w)$ 求极大值得到 w。

则回归模型为 $P(Y=1|x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)}$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x)}$$

多项逻辑斯蒂回归模型：

$$\begin{cases} P(Y=1|x)=\dfrac{\exp(w_1\cdot x)}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)} \\ P(Y=K|x)=\dfrac{1}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)} \end{cases}$$

6.2 最大熵模型

原理：在满足约束条件的模型集合中选取熵最大的模型

X服从均匀分布时，熵最大

最大熵模型的定义：

首先确定 $P(X,Y)$ 的经验分布和边缘分布 $P(X)$，$\tilde{P}(X)$

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

用特征函数 $f(x,y)$ 描述输入 x 和输出 y 之间的某一个事实

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

特征函数 $f(x,y)$ 关于经验分布 $\tilde{P}(X,Y)$ 的期望值，用 $E_{\tilde{P}}$ 表示

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

特征函数 $f(x,y)$ 关于模型 $P(Y|X)$ 与经验分布 $\tilde{P}(X)$ 的期望值，用 E_P 表示

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

假设满两个期望值相等：$E_P(f)=E_{\tilde{P}}(f)$ 为模型学习的约束条件

定义：$C=\{P\mid E_P(f_i)=E_{\tilde{P}}(f_i),\ i=1,2,\cdots,n\}$

定义在条件概率分布 $P(Y|X)$ 上的条件熵为

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

则模型集合 C 中条件熵 $H(P)$ 最大的模型称为最大熵模型

最大熵模型的学习：

满足约束条件的最优化问题：

$$\max_{P\in C} H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

$$s.t.\quad E_P(f_i)=E_{\tilde{P}}(f_i),\ i=1,2,\cdots,n$$

$$\sum_y P(y|x)=1$$

改写成最小值问题：

$$\min_{P\in C} -H(P)=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

转化为拉格朗日对偶问题：

① 定义拉格朗日函数 $L(P,w)$：

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

$$=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)+w_0\left(1-\sum_y P(y|x)\right)$$

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

② 原始问题：求对偶问题：

$$\min_{P\in C}\max_{w} L(P,w)\ \rightarrow\ \max_{w}\min_{P\in C} L(P,w)$$

③ 解对偶问题：

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

b. 将 $\Psi(w)$ 的解记为 $P_w=\arg\min_{P\in C}L(P,w)=P_w(y|x)$

c. 对 $L(P,w)$ 求 $P(y|x)$ 的偏导数

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

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

令 $=0$，$\tilde{P}(x)>0$ 的情况下，$P(y|x)=\frac{\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)}{\exp(1-w_0)}$

d. 归一化处理得：$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)$

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

最大熵模型 (x)

④ 解对偶问题外部的极大化问题：

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

极大似然估计：

证明对偶函数的极大化 $\Leftrightarrow$ 最大熵模型的极大似然估计

条件概率分布 $P(y|x)$ 的对数似然函数表示为：

$$L_{\tilde{P}}(P_w)=\log\prod_{x,y}P(y|x)^{\tilde{P}(x,y)}=\sum_{x,y}\tilde{P}(x,y)\log P(y|x)$$

当条件概率分布 $P(y|x)$ 是 x 时，对数似然函数 $L_{\tilde{P}}(P_w)$ 为

$$L_{\tilde{P}}(P_w)=\sum_{x,y}\tilde{P}(x,y)\log P(y|x)$$

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

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

又对偶函数 $\Psi(w)$：

$$\Psi(w)=\sum_{x,y}\tilde{P}(x)P_w(y|x)\log P_w(y|x)+\sum_{i=1}^{n}w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_x\tilde{P}(x)P_w(y|x)f_i(x,y)\right)$$

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

且 $\sum_y P(y|x)=1$，则得：$\Psi(w)=L_{\tilde{P}}(P_w)$

最大熵模型：

$$\begin{cases} P_w(y|x)=\dfrac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right) \\ Z_w(x)=\sum_y\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right) \end{cases}$$