课本 5.1,5.2

感知机：（参考：李航《统计学习方法》第二章）

公式解释：假设输入空间是$\mathcal{X} \subseteq R^n$，输出空间是$\mathcal{Y} \subseteq \{0,1\}$。输入$x \subseteq \mathcal{X}$表示实例的特征向量，对应于输入空间的点;输出$y \subseteq \mathcal{Y}$表示实例的类别。由输入空间到输出空间的如下函数：

$$f(x) = sgn(\boldsymbol{\omega^T x} + b)$$

式 1.1

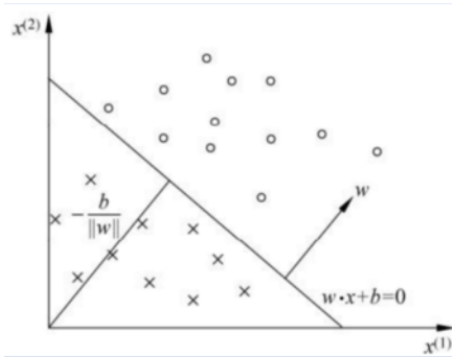
注：$\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots \omega_n]$，$\boldsymbol{x} = [x^1, \boldsymbol{x^2}, \ldots, \boldsymbol{x^n}]^T$ $x^i$代表特征向量 x 的第 i 个特征称为感知机。其中$\boldsymbol{\omega}$和 b 为感知机模型参数，sgn 是阶跃函数，即：

$$sgn(z) = \begin{cases} 1, z \geq 0 \\ 0, z < 0 \end{cases}$$

式 1.2

注意虽然 sgn 这里有≥，但是感知机超平面上是没有特征向量的!

几何解释：线性方程 $\boldsymbol{\omega^T x} + b = 0$对应于特征空间（输入空间）$R^n$ 中的一个超平面 S，其中$\boldsymbol{\omega}$是超平面的法向量，b 是超平面的截距。这个超平面将特征空间划分为两个部分。位于两边的点(特征向量）分别被分为正、负两类。因此超平面 S 称为分离超平面。如图所示：




**（超平面上方的点（$\boldsymbol{\omega^T x} + b > 0$）（○）被分为正类（1），**
**下方的点（$\boldsymbol{\omega^T x} + b < 0$）（×）被分为负类（0））**

学习策略：假设训练数据集是线性可分的，感知机学习的目标是求得一个能够将训练集正实例点和负实例点完全正确分开的超平面。为了找出这样的超平面 S，即确定感知机模型参数$\boldsymbol{\omega}$和 b，需要确定一个学习策略，即**定义损失函数并将损失函数极小化**。损失函数的一个自然选择是误分类点的总数（离散值）。但是，这样的损失函数不是参数 $\boldsymbol{\omega}$和 b 的连续可导函数，不易优化，所以**感知机采用的损失函数为误分类点到超平面的总距离。**
**感知机损失函数推导如下：**

输入空间$R^n$中点$\boldsymbol{x_0}$到超平面 S 的距离公式为：

$$\frac{\left|\boldsymbol{\omega^T x_0} + b\right|}{\|\boldsymbol{\omega}\|}$$

式 1.3

其中，$\|\boldsymbol{\omega}\|$表示向量$\boldsymbol{\omega}$的$L_2$范数，也即模长。若将 b 看成哑结点，也即合并到$\boldsymbol{\omega}$可得

$$\frac{\left|\boldsymbol{\hat{\omega}^T \hat{x}_0}\right|}{\|\boldsymbol{\hat{\omega}}\|}$$

式 1.4

注：$\boldsymbol{\hat{\omega}^T} = [\omega_1, \omega_2, \ldots \omega_n, \boldsymbol{b}]$，$\boldsymbol{\hat{x}_0} = [x_0^1, \boldsymbol{x_0^2}, \ldots, \boldsymbol{x_0^n}, \boldsymbol{1}]^T$

设误分类点集合为 M，那么所有误分类点到超平面 S 的总距离为

$$\sum_{\hat{x}_i \in M} \frac{|\hat{\omega}^T \hat{x}_i|}{\|\hat{\omega}\|}$$

式 1.5

**绝对值函数不易优化，想办法去掉它！**

对于任意误分类点$\hat{x}_i$，都有下表关系：

| 种类 | 实际位置 | $\hat{\omega}^T \hat{x}_i$ | 真实值$y_i$ | 误分类值$\hat{y}_i$ | $(\hat{y}_i - y_i)$ |
|---|---|---|---|---|---|
| ○ | 超平面上方 | >0 | 1 | 0 | 1 |
| × | 超平面下方 | <0 | 0 | 1 | -1 |

可见$\forall \hat{x}_i \in M$，恒有

$$(\hat{y}_i - y_i)\hat{\omega}^T \hat{x}_i > 0$$

式 1.6

于是所有误分类点到超平面 S 的总距离可改写为：

$$\sum_{\hat{x}_i \in M} \frac{(\hat{y}_i - y_i)\hat{\omega}^T \hat{x}_i}{\|\hat{\omega}\|}$$

式 1.7

**式 7 和式 5 等价，利用误分类关系巧妙地去掉了绝对值。**

**理想情况下就是所有点正确分类，没有误分类点。对应式 7 就是希望它尽可能小，为 0 最好。对于分式而言分母不可为 0，所以只要求极小化分子即可。**

**所以感知机的学习损失函数为：**

$$L(\hat{\omega}) = \sum_{\hat{x}_i \in M} (\hat{y}_i - y_i)\hat{\omega}^T \hat{x}_i$$

式 1.8

显然，损失函数$L(\hat{\omega})$是非负的。如果没有误分类点，损失函数值是 0。而且，误分类点越少，误分类点离超平面越近，损失函数值就越小，在误分类时是参数$\hat{\omega}$的线性函数，在正确分类时是 0。因此，给定训练数据集，损失函数$L(\hat{\omega})$是$\hat{\omega}$的连续可导函数。

学习算法：**(如何极小化损失函数式 8？)**感知机学习算法是对以下最优化问题的算法，给定训练数据集:

$$T = \{(\hat{x}_1, y_1), (\hat{x}_2, y_2), \dots, (\hat{x}_n, y_n)\}$$

其中$\hat{x}_i \in R^{n+1}, y \in \{0,1\}$ 求参数$\hat{\omega}$使其为以下损失函数极小化问题的解

$$L(\hat{\omega}) = \sum_{\hat{x}_i \in M} (\hat{y}_i - y_i)\hat{\omega}^T \hat{x}_i$$

式 1.8

其中 M 为误分类点的集合

感知机学习算法是误分类驱动的，具体**采用随机梯度下降法**。

首先，任意选取一个超平面$\hat{\omega}_0{}^T \hat{x}$=0 用梯度下降法不断地极小化损失函数$L(\hat{\omega})$，极小化过程中不是一次使 M 中所有误分类点的梯度下降，而是一次随机选取一个误分类点使其梯度下降。

已知损失函数的梯度为:

$$\nabla \boldsymbol{L}(\hat{\boldsymbol{\omega}}) = \frac{\partial L(\hat{\boldsymbol{\omega}})}{\partial \hat{\boldsymbol{\omega}}} = \frac{\partial}{\partial \hat{\boldsymbol{\omega}}}\left[\sum_{\hat{\boldsymbol{x}}_i \in \boldsymbol{M}} (\hat{y}_i - y_i)\hat{\boldsymbol{x}}_{\boldsymbol{i}}\right]$$

$$= \sum_{\hat{\boldsymbol{x}}_i \in \boldsymbol{M}}\left[(\hat{y}_i - y_i)\frac{\partial}{\partial \hat{\boldsymbol{\omega}}}\hat{\boldsymbol{\omega}}^{\boldsymbol{T}}\hat{\boldsymbol{x}}_{\boldsymbol{i}}\right]$$

式 1.9

由矩阵微分公式$\frac{\partial \boldsymbol{x}^{\boldsymbol{T}} a}{\partial \boldsymbol{x}} = \boldsymbol{a}$可得,

$$\nabla \boldsymbol{L}(\hat{\boldsymbol{\omega}}) = \frac{\partial L(\hat{\boldsymbol{\omega}})}{\partial \hat{\boldsymbol{\omega}}} = \sum_{\hat{\boldsymbol{x}}_i \in \boldsymbol{M}} (\hat{y}_i - y_i)\hat{\boldsymbol{x}}_{\boldsymbol{i}}$$

式 1.10

由于参数$\hat{\boldsymbol{\omega}}$的更新过程为

$$\hat{\boldsymbol{\omega}} \leftarrow \hat{\boldsymbol{\omega}} + \Delta\hat{\boldsymbol{\omega}}$$

式 1.11

$\because \Delta\hat{\boldsymbol{\omega}} = -\eta\nabla \boldsymbol{L}(\hat{\boldsymbol{\omega}})\ \ \eta\epsilon(0,1)$ （沿负梯度方向以学习率$\eta$对$\hat{\boldsymbol{\omega}}$更新）

$\therefore \hat{\boldsymbol{\omega}} \leftarrow \hat{\boldsymbol{\omega}} - \eta\nabla \boldsymbol{L}(\hat{\boldsymbol{\omega}})$

随机选取一个误分类点$\hat{\boldsymbol{x}}_{\boldsymbol{i}}$进行梯度下降，也即指式 10 在代入时没有求和号，故取$\nabla \boldsymbol{L}(\hat{\boldsymbol{\omega}}) = (\hat{y}_i - y_i)\hat{\boldsymbol{x}}_{\boldsymbol{i}}$代入上式

得 $\hat{\boldsymbol{\omega}} \leftarrow \hat{\boldsymbol{\omega}} - \eta(\hat{y}_i - y_i)\,\hat{\boldsymbol{x}}_{\boldsymbol{i}} = \hat{\boldsymbol{\omega}} + \eta(y_i - \hat{y}_i)\,\hat{\boldsymbol{x}}_{\boldsymbol{i}}$

其中定义

$$\Delta\hat{\boldsymbol{\omega}} = \eta(y_i - \hat{y}_i)\,\hat{\boldsymbol{x}}_{\boldsymbol{i}}$$

式 1.12

由此可见，式 1.11，式 1.12，对应课本 5.1,5.2

课本 5.10 5.11 5.12 5.13 5.14 5.15

**标准 BP 算法：**

给定一个训练样本$(x_k,y_k)$，假设模型输出为$\hat{y}_k=(\hat{y}_1^k,\hat{y}_2^k,\ldots,\hat{y}_l^k)$。则**均方误差**为：

$$E_k=\frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k-y_j^k\right)^2$$

式 2.1

注：$\hat{y}_j^k$为第 k 个训练样本$x_k$,经神经网络输出第 j 类（共 l 类）的概率值。$y_j^k$为训练样本$x_k$,的第 j 类（共 l 类）的真实值。如三分类问题（l=3）$\hat{y}_k=\left(\hat{y}_1^k,\hat{y}_2^k,\hat{y}_3^k\right)=(0.9,0.05,0.05)$ $y_k=(y_1^k,y_2^k,y_3^k)=(1,0,0)$

如果按照梯度下降法更新模型参数，那么**各个参数的更新公式**为：

$$\omega_{hj}\leftarrow\omega_{hj}+\Delta\omega_{hj}=\ \omega_{hj}-\eta\frac{\partial E_k}{\partial\omega_{hj}}$$

式 2.2

$$\theta_j\leftarrow\theta_j+\Delta\theta_j=\ \theta_j-\eta\frac{\partial E_k}{\partial\theta_j}$$

式 2.3

$$v_{ih}\leftarrow v_{ih}+\Delta v_{ih}=\ v_{ih}-\eta\frac{\partial E_k}{\partial v_{ih}}$$

式 2.4

$$\gamma_h\leftarrow\gamma_h+\Delta\gamma_h=\ v_{ih}-\eta\frac{\partial E_k}{\partial\gamma_h}$$

式 2.5

显然，**只要能求得四个偏导就可以确定更新公式了**。

1 求$\frac{\partial E_k}{\partial\omega_{hj}}$

已知$\boldsymbol{E_k}$**和**$\boldsymbol{\omega_{hj}}$**的函数链式关系**为：

$$E_k=\frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k-y_j^k\right)^2$$

式 2.1

$$\hat{y}_j^k=f(\beta_j-\theta_j)\qquad\text{f 为 sigmoid 函数}$$

式 2.6

$$\beta_j=\sum_{h=1}^{q}\omega_{hj}b_h$$

式 2.7

所以

$$\frac{\partial E_k}{\partial \omega_{hj}} = \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial \omega_{hj}}$$

1.1 求$\frac{\partial E_k}{\partial \hat{y}_j^k}$

$$\frac{\partial E_k}{\partial \hat{y}_j^k} = \frac{\partial[\frac{1}{2}\sum_{j=1}^{l}(\hat{y}_j^k - y_j^k)^2]}{\partial \hat{y}_j^k} = \frac{\frac{1}{2}\partial[(\hat{y}_1^k - y_1^k)^2 + \ldots + (\hat{y}_j^k - y_j^k)^2 + \ldots + (\hat{y}_l^k - y_l^k)^2]}{\partial \hat{y}_j^k}$$

$$= \frac{1}{2} * 2 * (\hat{y}_j^k - y_j^k) * 1 = (\hat{y}_j^k - y_j^k)$$

式 2.8

1.2 求$\frac{\partial \hat{y}_j^k}{\partial \beta_j}$

$$\frac{\partial \hat{y}_j^k}{\partial \beta_j} = \frac{\partial f(\beta_j - \theta_j)}{\partial \beta_j} = f'(\beta_j - \theta_j)$$

式 2.9

由于$f(x)$为 sigmoid 函数，有特殊性质：$f'(x) = f(x)(1 - f(x))$

式 2.10

式 2.6，式 2.10 代入式 2.9 可得：

$$\frac{\partial \hat{y}_j^k}{\partial \beta_j} = f(\beta_j - \theta_j) * [1 - f(\beta_j - \theta_j)] = \hat{y}_j^k(1 - \hat{y}_j^k)$$

式 2.11

1.3 求$\frac{\partial \beta_j}{\partial \omega_{hj}}$

$$\frac{\partial \beta_j}{\partial \omega_{hj}} = \frac{\partial(\sum_{h=1}^{q}\omega_{hj}b_h)}{\partial \omega_{hj}} = b_h$$

式 2.12

由式 2.8， 2.11 令

$$g_j = -\frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j} = -(\hat{y}_j^k - y_j^k)\hat{y}_j^k(1 - \hat{y}_j^k)$$

式 2.13 即为课本式 5.10

由式 2.2

$$\Delta\omega_{hj} = -\eta\frac{\partial E_k}{\partial \omega_{hj}} = -\eta\frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial \omega_{hj}} = \eta g_j b_h$$

式 2.14 即为课本式 5.11

2 求$\frac{\partial E_k}{\partial \theta_j}$

已知$E_k$和$\theta_j$的函数链式关系为：

$$E_k = \frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k - y_j^k\right)^2$$

式 2.1

$$\hat{y}_j^k = \hat{y}_j^k = f(\beta_j - \theta_j)$$

式 2.6

根据链式求导法则可得

$$\frac{\partial E_k}{\partial \theta_j} = \frac{\partial E_k}{\partial\, \hat{y}_j^k}\frac{\partial\, \hat{y}_j^k}{\partial \theta_j}$$

式 2.15

由式 2.8 2.6 2.10 2.11 代入 2.15 得：

$$\frac{\partial E_k}{\partial \theta_j} = \left(\hat{y}_j^k - y_j^k\right)\frac{\partial\, \hat{y}_j^k}{\partial \theta_j} = \left(\hat{y}_j^k - y_j^k\right)\frac{\partial f(\beta_j - \theta_j)}{\partial \theta_j} = \left(\hat{y}_j^k - y_j^k\right)f'(\beta_j - \theta_j) * (-1)$$
$$= \left(y_j^k - \hat{y}_j^k\right)f'(\beta_j - \theta_j) = \left(y_j^k - \hat{y}_j^k\right)\hat{y}_j^k\left(1 - \hat{y}_j^k\right)$$

式 2.16

所以

$$\Delta\theta_j = -\eta\frac{\partial E_k}{\partial \theta_j} = -\eta\left(y_j^k - \hat{y}_j^k\right)\hat{y}_j^k\left(1 - \hat{y}_j^k\right) = \eta g_j$$

式 2.17 即为课本上式 5.12

3、求$\frac{\partial E_k}{\partial v_{ih}}$

已知$E_k$和$v_{ih}$的函数链式关系为：

$$E_k = \frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k - y_j^k\right)^2$$

式 2.1

$$\hat{y}_j^k = f(\beta_j - \theta_j) \qquad \text{f 为 sigmoid 函数}$$

式 2.6

$$\beta_j = \sum_{h=1}^{q}\omega_{hj}b_h$$

式 2.7

$$b_h = f(\alpha_h - \gamma_h)$$

式 2.18

$$\alpha_h = \sum_{i=1}^{d}v_{ih}x_i$$

式 2.19

根据链式求导法则可得

$$\frac{\partial E_k}{\partial v_{ih}} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \frac{\partial \hat{y}_j^k}{\partial \beta_j} \frac{\partial \beta_j}{\partial b_h} \frac{\partial b_h}{\partial \alpha_\hbar} \frac{\partial \alpha_\hbar}{\partial v_{ih}}$$

式 2.20

式 2.20 的求和符号产生的原因在于：式 2.19 中 1 个$v_{ih}$对应一个$\alpha_\hbar$，式 2.18 中，1 个$\alpha_\hbar$对应一个$b_h$，可是式 2.7 中，l 个$\beta_j$里每个$\beta_j$都包含一个$b_h$。（带$v_{23}$进去 2.19 往上算到 2.7 就知道了）。

3.1 求$\frac{\partial \beta_j}{\partial b_h}$

由 2.7

$$\frac{\partial \beta_j}{\partial b_h} = \frac{\partial \left( \sum_{h=1}^{q} \omega_{\hbar j} b_h \right)}{\partial b_h} = \omega_{\hbar j}$$

式 2.21

3.2 求$\frac{\partial b_h}{\partial \alpha_\hbar}$

由式 2.18

$$\frac{\partial b_h}{\partial \alpha_\hbar} = \frac{\partial f\left(\alpha_\hbar - \gamma_h\right)}{\partial \alpha_\hbar} = f'\left(\alpha_\hbar - \gamma_h\right) * 1 = f\left(\alpha_\hbar - \gamma_h\right) * \left[1 - f\left(\alpha_\hbar - \gamma_h\right)\right] = b_h * (1 - b_h)$$

式 2.22

3.3 求$\frac{\partial \alpha_\hbar}{\partial v_{ih}}$

由式 2.19

$$\frac{\partial \alpha_\hbar}{\partial v_{ih}} = \frac{\partial (\sum_{i=1}^{d} v_{ih} x_i)}{\partial v_{ih}} = x_i$$

式 2.23

令

$$e_h = -\frac{\partial E_k}{\partial \alpha_\hbar} = -\sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \frac{\partial \hat{y}_j^k}{\partial \beta_j} \frac{\partial \beta_j}{\partial b_h} \frac{\partial b_h}{\partial \alpha_\hbar} = b_h * (1 - b_h) \sum_{j=1}^{l} \omega_{\hbar j}\, g_j$$

式 2.24 即为课本式 5.15

所以

$$\Delta v_{ih} = -\eta \frac{\partial E_k}{\partial v_{ih}} = -\eta \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \frac{\partial \hat{y}_j^k}{\partial \beta_j} \frac{\partial \beta_j}{\partial b_h} \frac{\partial b_h}{\partial \alpha_\hbar} \frac{\partial \alpha_\hbar}{\partial v_{ih}} = \eta e_h x_i$$

式 2.25 即为课本式 5.13

4 求$\frac{\partial E_k}{\partial \gamma_h}$

已知$E_k$和$\gamma_h$的函数链式关系为：

$$E_k = \frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k - y_j^k\right)^2$$

式 2.1

$$\hat{y}_j^k = f(\beta_j - \theta_j) \qquad \text{f 为 sigmoid 函数}$$

式 2.6

$$\beta_j = \sum_{h=1}^{q} \omega_{hj} b_h$$

式 2.7

$$b_h = f(\alpha_h - \gamma_h)$$

式 2.18

根据链式求导法则可得

$$\frac{\partial E_k}{\partial \gamma_h} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial b_h}\frac{\partial b_h}{\partial \gamma_h}$$

式 2.26

由式 2.18

$$\frac{\partial E_k}{\partial \gamma_h} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial b_h}\frac{\partial b_h}{\partial \gamma_h} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial b_h}\frac{\partial f(\alpha_h - \gamma_h)}{\partial \gamma_h}$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial b_h} f'\left(\alpha_h - \gamma_h\right) * (-1)$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial b_h} f(\alpha_h - \gamma_h) * \left(1 - f(\alpha_h - \gamma_h)\right) * (-1)$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\frac{\partial \beta_j}{\partial b_h} b_h * \left(1 - b_h\right) * (-1)$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k}\frac{\partial \hat{y}_j^k}{\partial \beta_j}\omega_{hj} b_h * \left(1 - b_h\right) * (-1) = \sum_{j=1}^{l} g_j \omega_{hj} b_h * \left(1 - b_h\right) = e_h$$

式 2.27

所以

$$\Delta\gamma_h = -\eta\frac{\partial E_k}{\partial \gamma_h} = -\eta e_h$$

式 2.28 即为课本式 5.14