线性分类

频率派 → 统计机器学习.

贝叶斯派 → 概率图模型.

线性回归 $f(w,b)=w^Tx+b$, $x\in\mathbb{R}^p$

- ①线性
  - 属性：x是个向量，f关于x是线性的.
  - 全局：w^Tx+b是线性组合，直接输出.
  - 系数：f关于w是线性的.
- ②全局性：在整个特征空间拟合一条直线，特殊的.
- ③数据未加工：直接使用数据.

线性回归是统计机器学习的基本模型，其它模型通过打破其一个或多个特点形成了统计机器学习的架构.

若打破属性线性，如多项式回归，其属性可能是多次的，如x^2

全局线性：如线性分类，将f作为激活函数的输入，感知机.

系数线性：神经网络，感知机

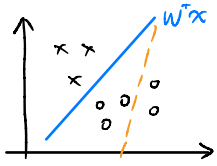
若打破全局性：对样本空间分割，如线性样条回归，决策树.

若打破数据未加工：PCA，流形.

线性回归 $\xrightarrow[\text{降维}]{\text{激活函数}}$ 线性分类.

线性分类
- 硬分类 $y\in\{0,1\}$
  - LDA
  - 感知机.
- 软分类 $y\in[0,1]$
  - 生成式：高斯判别分析.
  - 判别式：LR

感知机 (Perceptron)



思想：错误驱动

模型：$f(x)=\text{sign}(w^Tx)$，$x\in\mathbb{R}^p$，$w\in\mathbb{R}^p$，样本个数为N.

指示函数，$\text{sign}(a)=\begin{cases}+1, & a\geq 0\\ -1, & a<0\end{cases}$

整体思想：给定初始w（图中橙线），会获得被分类错误的样本集合D，依次取D中的样本，调整w，使得w对当前取得的样本分类正确.

策略：loss function

$$L(w)=\sum_{i=1}^{N} I\{y_i w^Tx<0\}$$

其中I为指示函数，当I中条件成立时，$I=1$，否则$I=0$.

但当w变化Δw时，I可能从0变为1，它是不可导的，因此这个loss不合适.

当正确分类时，对任意样本x_i，有：

$$\begin{cases}w^Tx_i>0, & y_i=+1\\ w^Tx_i<0, & y_i=-1\end{cases}\Rightarrow y_i w^Tx_i>0.$$

当错误分类时，对任意样本x_i，有：

$$\begin{cases}w^Tx_i>0, & y_i=-1\\ w^Tx_i<0, & y_i=+1\end{cases}\Rightarrow y_i w^Tx_i<0.$$

$$L(w)=\sum_{x_i\in D}-y_i w^Tx_i$$

由于每个错误分类的点有$y_i w^Tx_i$都小于0，因此$-y_i w^Tx_i$大于0，此值可作为损失值，将所有误分类点有的$-y_i w^Tx_i$求和作为损失函数即可.

算法：SGD（随机梯度下降）

$$w^{(t+1)}\leftarrow w^{(t)}-\lambda\nabla L$$

线性判别分析（LDA）

设：$X=(x_1,x_2,\cdots,x_N)^T=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}_{N\times p}$

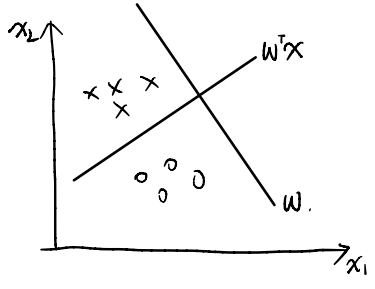
$Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_N\end{pmatrix}_{N\times 1}$

$\{(x_i,y_i)\}_{i=1}^N$，$x_i\in\mathbb{R}^p$，$y_i\in\{+1,-1\}$

$x_{C_1}=\{x_i\mid y_i=+1\}$，$x_{C_2}=\{x_i\mid y_i=-1\}$

$|x_{C_1}|=N_1$　$|x_{C_2}|=N_2$　$N_1+N_2=N$.

LDA 的思想：类内小，类间大.



限定 $\|w\|=1$.

样本点在 w 上的投影：$z_i=w^Tx_i$

样本投影的均值：$\bar{z}=\frac{1}{N}\sum_{i=1}^N z_i=\frac{1}{N}\sum_{i=1}^N w^Tx_i$

z 的协方差矩阵：$S_z=\frac{1}{N}\sum_{i=1}^N(z_i-\bar{z})(z_i-\bar{z})^T$

$=\frac{1}{N}\sum_{i=1}^N(w^Tx_i-\bar{z})(w^Tx_i-\bar{z})^T$

C_1：$\bar{z}_1=\frac{1}{N_1}\sum_{i=1}^{N_1}w^Tx_i$

$S_1=\frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i-\bar{z}_1)(w^Tx_i-\bar{z}_1)^T$

C_2：$\bar{z}_2=\frac{1}{N_2}\sum_{i=1}^{N_2}w^Tx_i$

$S_2=\frac{1}{N_2}\sum_{i=1}^{N_2}(w^Tx_i-\bar{z}_2)(w^Tx_i-\bar{z}_2)^T$

类间距离：$(\bar{z}_1-\bar{z}_2)^2$

类内距离：S_1+S_2

目标函数：$J(w)=\frac{(\bar{z}_1-\bar{z}_2)^2}{S_1+S_2}$，则 $\hat{w}=\arg\max_w J(w)$

$$J(w)=\frac{(\bar{z}_1-\bar{z}_2)^2}{S_1+S_2}$$

分子$=\left(\frac{1}{N_1}\sum_{i=1}^{N_1}w^Tx_i-\frac{1}{N_2}\sum_{i=1}^{N_2}w^Tx_i\right)^2=\left(w^T\left(\frac{1}{N_1}\sum_{i=1}^{N_1}x_i-\frac{1}{N_2}\sum_{i=1}^{N_2}x_i\right)\right)^2=\left(w^T(\bar{x}_{C_1}-\bar{x}_{C_2})\right)^2$

分母 $= S_1 + S_2$

$$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_j)(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_j)^T$$

$$= \frac{1}{N_1}\sum_{i=1}^{N_1} w^T(x_i - \bar{x}_{c_1})(x_i - \bar{x}_{c_1})^T w$$

$$= w^T \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N_1}(x_i - \bar{x}_{c_1})(x_i - \bar{x}_{c_1})^T\right]}_{S_{c_1}} w$$

$$= w^T S_{c_1} w$$

故分母 $= w^T S_{c_1} w + w^T S_{c_2} w$

$$= w^T (S_{c_1} + S_{c_2}) w$$

故 $J(w) = \dfrac{w^T(\bar{x}_{c_1} - \bar{x}_{c_2})(\bar{x}_{c_1} - \bar{x}_{c_2})^T w}{w^T(S_{c_1} - S_{c_2})w}$

模型求解：$J(w) = \dfrac{w^T S_b w}{w^T S_w w}$，其中 S_b：类间方差，S_w：类内方差.

$$J(w) = w^T S_b w (w^T S_w w)^{-1}$$

令 $\dfrac{\partial J(w)}{\partial w} = 2S_b w (w^T S_w w)^{-1} + w^T S_b w \cdot (-1)\cdot(w^T S_w w)^{-2}\cdot 2S_w \cdot w = 0$

即 $S_b w (w^T S_w w) - w^T S_b w S_w w = 0$

$$\underbrace{w^T S_b w}_{1\times 1 \in R} S_w w = S_b w \underbrace{(w^T S_w w)}_{1\times p \cdot p\times p \cdot p\times 1 = 1\times 1 \in R}$$

则 $S_w \cdot w = \dfrac{w^T S_w w}{w^T S_b w} \cdot S_b w$

$$w = \frac{w^T S_w w}{w^T S_b w} \cdot S_w^{-1} \cdot S_b \cdot w \propto S_w^{-1} \cdot \underbrace{S_b \cdot w}_{(\bar{x}_{c_1} - \bar{x}_{c_2})\underbrace{(\bar{x}_{c_1} - \bar{x}_{c_2})^T w}_{1\times 1}}$$

$$\propto S_w^{-1}(\bar{x}_{c_1} - \bar{x}_{c_2})$$

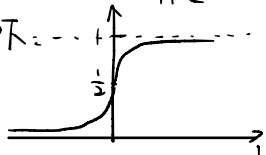
当类别有 K 个时，最多降维到 K-1 维.

Logistic Regression

判别模型直接对 $P(Y|X)$ 进行建模，应用 MLE.

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

其图象如下：



$z \to +\infty$, $\sigma(z) \to 1$

$z = 0$, $\sigma(z) = \frac{1}{2}$

$z \to -\infty$, $\sigma(z) \to 0$

$$\begin{cases} p_1 = P(y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}, & y=1 \\ p_0 = P(y=0|x) = 1 - \frac{1}{1+e^{-w^T x}} = \frac{e^{-w^T x}}{1+e^{-w^T x}}, & y=0 \end{cases}$$

$$\Rightarrow P(y|x) = p_1^{y} p_0^{1-y}$$

$$\begin{aligned} \text{MLE}: \hat{w} &= \arg\max_w \log P(Y|X) \\ &= \arg\max_w \log \prod_{i=1}^{N} P(y_i|x_i) \\ &= \arg\max_w \sum_{i=1}^{N} \log P(y_i|x_i) \\ &= \arg\max_w \sum_{i=1}^{N} (y_i \log p_1 + (1-y_i) \log p_0) \end{aligned}$$

$-$ cross entropy

MLE $\xrightarrow{(\max)}$ loss function $\xrightarrow{(\min)}$ (min cross entropy)

高斯判别分析（Gaussian Discriminant Analysis）

Data：$\{(x_i, y_i)\}_{i=1}^N$，$x_i \in \mathbb{R}^p$，$y_i \in \{0,1\}$

前面的线性分类模型都是判别模型，直接对 $P(Y|X)$ 进行建模，即求 $\hat{y} = \arg\max_{y \in \{0,1\}} P(y|x)$

而高斯判别模型是生成模型，应用贝叶斯定理：$P(Y|X) = \frac{P(X|Y)\cdot P(Y)}{P(X)} \propto P(X|Y)\cdot P(Y)$

$= P(X,Y)$，所以生成模型是对联合概率进行建模

即求 $\hat{y} = \arg\max_{y \in \{0,1\}} P(y)\cdot P(x|y)$

GDA：

对于 y，可将 y 看作伯努利分布，$y \sim \text{Bernoulli}(\phi)$

y	1	0
	ϕ	$1-\phi$

$P(y) = \phi^y \cdot (1-\phi)^{1-y}$

令 $x|y=1 \sim N(\mu_1, \Sigma)$，$x|y=0 \sim N(\mu_2, \Sigma)$ $\Rightarrow N(\mu_1,\Sigma)^y \cdot N(\mu_2,\Sigma)^{1-y}$

log-likelihood：$\theta = (\mu_1, \mu_2, \Sigma, \phi)$

$$
\begin{aligned}
l(\theta) &= \log \prod_{i=1}^N P(x_i, y_i) \\
&= \sum_{i=1}^N \log\left[P(x_i|y_i)\cdot P(y_i)\right] \\
&= \sum_{i=1}^N \left[\log P(x_i|y_i) + \log P(y_i)\right] \\
&= \sum_{i=1}^N \left[\log N(\mu_1,\Sigma)^{y_i}\cdot N(\mu_2,\Sigma)^{1-y_i} + \log \phi^{y_i}\cdot(1-\phi)^{1-y_i}\right] \\
&= \sum_{i=1}^N \left[\underbrace{\log N(\mu_1,\Sigma)^{y_i}}_{①} + \underbrace{\log N(\mu_2,\Sigma)^{1-y_i}}_{②} + \underbrace{\log \phi^{y_i}\cdot(1-\phi)^{1-y_i}}_{③}\right]
\end{aligned}
$$

$\hat{\theta} = \arg\max_\theta l(\theta)$

模型求解

求 ϕ：

$$
\begin{aligned}
③ &= \sum_{i=1}^N y_i \log\phi + (1-y_i)\log(1-\phi) \\
\frac{\partial ③}{\partial \phi} &= \sum_{i=1}^N y_i \cdot \frac{1}{\phi} - (1-y_i)\frac{1}{1-\phi} = 0 \\
\Rightarrow \sum_{i=1}^N y_i(1-\phi) - (1-y_i)\phi &= 0 \\
\sum_{i=1}^N y_i - y_i\phi - \phi + y_i\phi &= 0 \\
\sum_{i=1}^N (y_i - \phi) &= 0 \\
\sum_{i=1}^N y_i - N\phi &= 0
\end{aligned}
$$

$\therefore \boxed{\hat{\phi} = \frac{1}{N}\sum_{i=1}^N y_i}$

设 $y=1$ 的样本为 N_1 个，$y=0$ 的样本 N_2 个，$N=N_1+N_2$

则 $\hat{\phi}=\frac{1}{N}\sum_{i=1}^{N} y_i=\frac{N_1}{N}$

求 μ_1 (μ_2)：

$$①=\sum_{i=1}^{N}\log N(\mu_1,\Sigma)^{y_i}$$

$$=\sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right)$$

$$\mu_1=\arg\max_{\mu_1} ①=\arg\max_{\mu_1}\sum_{i=1}^{N} y_i\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu_1)\right)$$

$$\Delta=\sum_{i=1}^{N} y_i\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)=-\frac{1}{2}\sum_{i=1}^{N} y_i\left(x_i^T\Sigma^{-1}-\mu_1^T\Sigma^{-1}\right)(x_i-\mu_1)$$

$$=-\frac{1}{2}\sum_{i=1}^{N} y_i\left(x_i^T\Sigma^{-1}x_i-x_i^T\Sigma^{-1}\mu_1-\mu_1^T\Sigma^{-1}x_i+\mu_1^T\Sigma^{-1}\mu_1\right)$$

$$=-\frac{1}{2}\sum_{i=1}^{N} y_i\left(x_i^T\Sigma^{-1}x_i-2\mu_1^T\Sigma^{-1}x_i+\mu_1^T\Sigma^{-1}\mu_1\right)$$

$$\frac{\partial\Delta}{\partial\mu_1}=-\frac{1}{2}\sum_{i=1}^{N} y_i\left(-2\Sigma^{-1}x_i+2\Sigma^{-1}\mu_1\right)=0$$

$$\sum_{i=1}^{N} y_i\left(\Sigma^{-1}\mu_1-\Sigma^{-1}x_i\right)=0$$

$$\sum_{i=1}^{N} y_i(\mu_1-x_i)=0$$

则 $\hat{\mu}_1=\frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i}=\frac{\sum_{i=1}^{N} y_i x_i}{N_1}$　　μ_2 同理.

求 Σ：

设 $y_i=1$ 的样本点集合为 C_1，$y_i=0$ 的样本点集合为 C_2

$$\hat{\Sigma}=\arg\max_{\Sigma} ①+②$$

$$①+②=\sum_{x_i\in C_1}\log N(\mu_1,\Sigma)+\sum_{x_i\in C_2}\log N(\mu_2,\Sigma)$$

由 $\sum_{i=1}^{N}\log N(\mu,\Sigma)=\sum_{i=1}^{N}\left[\log\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\}\right]$

$$=\sum_{i=1}^{N}\left[\log\frac{1}{(2\pi)^{\frac{p}{2}}}+\log|\Sigma|^{-\frac{1}{2}}+\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\right]$$

$$=\sum_{i=1}^{N}\left[C-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]$$

$$=C-\frac{1}{2}N\log|\Sigma|-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

实数的迹等于它本身.

$$\sum_{i=1}^{N} tr\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

一个系数的方差矩阵.

$$=\sum_{i=1}^{N} tr\left((x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)$$

$$=tr\left(\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)$$

$=N\cdot tr(S\cdot\Sigma^{-1})$

$$= -\frac{1}{2}N \cdot \log|\Sigma| - \frac{1}{2}N \cdot tr(S \cdot \Sigma^{-1}) + C$$

取 ① + ② $= -\frac{1}{2}N_1 \log|\Sigma| - \frac{1}{2}N_1 tr(S_1\Sigma^{-1}) - \frac{1}{2}N_2 \log|\Sigma| - \frac{1}{2}N_2 tr(S_2\Sigma^{-1}) + C$

$= -\frac{1}{2}[N \log|\Sigma| + N_1 \cdot tr(S_1\Sigma^{-1}) + N_2 \cdot tr(S_2\Sigma^{-1})] + C$

$\frac{\partial tr(AB)}{\partial A} = B^T$

$\frac{\partial |A|}{\partial A} = |A| \cdot A^{-1}$

$$\frac{\partial ① + ②}{\partial \Sigma} = -\frac{1}{2}\left(N\Sigma^{-1} + N_1 \cdot \frac{\partial tr(S_1\Sigma^{-1})}{\partial \Sigma^{-1}} \cdot \frac{\partial \Sigma^{-1}}{\partial \Sigma} + N_2 \cdot \frac{\partial tr(S_2\Sigma^{-1})}{\partial \Sigma^{-1}} \cdot \frac{\partial \Sigma^{-1}}{\partial \Sigma}\right)$$

$$= -\frac{1}{2}(N \cdot \Sigma^{-1} - N_1 S_1^T \Sigma^{-2} - N_2 S_2^T \Sigma^{-2})$$

$$= -\frac{1}{2}(N \cdot \Sigma^{-1} - N_1 S_1 \Sigma^{-2} - N_2 S_2 \Sigma^{-2}) = 0$$

取 $N\Sigma - N_1 S_1 - N_2 S_2 = 0$

$$\hat{\Sigma} = \frac{N_1 S_1 + N_2 S_2}{N}$$