统计机器学习
Statistical Machine Learning

魏莱

上海海事大学信息工程学院

2019 年 3 月 25 日

第三章：线性模型的概率解释及线性判别分析

线性模型 $f(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$，假设现在有 $\mathbf{x}_i$，那么 $f(\mathbf{x}_i)$ 为 $\mathbf{x}_i$ 的一个预测。令 y_i 表示 $\mathbf{x}_i$ 的真实输出，则可以定义两者误差为

$$\epsilon_i = y_i - f(\mathbf{x}_i) = y_i - \mathbf{w}^t\mathbf{x}_i \tag{1}$$

在通常情况下，我们可以假设误差 $\epsilon_i, i = 1, \cdots, n$（$n$ 为样本数量）独立同分布 (independently and identically distributed，IID)，并且分布满足正太分布。

因此，有 $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$。由于 ϵ 分布随机，因此可以令 $\mu = 0$。再观察，由于 $y_i = \mathbf{w}^t \mathbf{x}_i + \epsilon_i$，根据正太分布性质，可知 $y_i \sim \mathcal{N}(\mathbf{w}^t \mathbf{x}_i, \sigma^2)$。因此有：

$$p(y_i|\mathbf{x}_i; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mathbf{w}^t \mathbf{x}_i)^2}{2\sigma^2}\right) \tag{2}$$

对于所有数据样本 $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$。

根据 (2)，假设其中 $\mathbf{y} = (y_1, \cdots, y_n)^t$，$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$，则联合概率密度可以表示成 $p(\mathbf{y}|\mathbf{X}; \mathbf{w})$。由于样本独立同分布（合理假设），因此，有

$$p(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mathbf{w}^t\mathbf{x}_i)^2}{2\sigma^2}\right) \tag{3}$$

现在思考，其中 $\mathbf{w}$ 的选择问题，什么样的 $\mathbf{w}$ 才是最优的？

由于我们获取了当前数据集，因此可以想象，当前数据样本出现的概率应该是比较大的。因此，我们应该选择一个 $\mathbf{w}$，使得当前数据样本出现的概率最大——最大似然 (maximum likelihood)。于是

$$\begin{aligned}\mathbf{w}_{opt} &= \arg\max p(\mathbf{y}|\mathbf{X};\mathbf{w}) = \arg\max L(\mathbf{w}) \\ &= \arg\max \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mathbf{w}^t\mathbf{x}_i)^2}{2\sigma^2}\right)\end{aligned} \tag{4}$$

这里 $L(\mathbf{w})$ 称为**似然函数 (likelihood function)**。然后请问如何求解 $\mathbf{w}$？

由于上式带有对数运算及联乘运算，为了求解方便，我们可以定义**对数似然函数**:

$$
\begin{aligned}
l(\mathbf{w}) &= \log L(\mathbf{w}) \\
&= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mathbf{w}^t\mathbf{x}_i)^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mathbf{w}^t\mathbf{x}_i)^2}{2\sigma^2}\right)\right) \\
&= n\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^t\mathbf{x}_i)^2
\end{aligned} \tag{5}
$$

由于对数函数的单调性，最大化 $l(\mathbf{w})$ 等价于最大化 $L(\mathbf{w})$. 由此可得，$\max_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^t\mathbf{x}_i)^2 = \min_{\mathbf{w}} E(\mathbf{w})$。

2. 正则线性回归的概率解释

在上述讨论中，我们认为 $\mathbf{w}$ 是一个固定的待求参数。如果假设 $\mathbf{w}$ 也是一个随机变量，那么我们应该如何求解 $\mathbf{w}$？

我们首先假设 $\mathbf{w}$ 也满足正太分布，同样由于随机性，可以得到 $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$。则，

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= p(\mathbf{y}|\mathbf{X}; \mathbf{w})p(\mathbf{w}) \\ &= \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(\frac{-\mathbf{w}^t\mathbf{w}}{2\lambda^{-1}}\right) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mathbf{w}^t\mathbf{x}_i)^2}{2\sigma^2}\right) \end{aligned} \tag{6}$$

等式第一行成立，可以这样解释，联合概率密度 $p(\mathbf{X}, \mathbf{w})$ 等于条件概率密度 $p(\mathbf{X}|\mathbf{w})$ 乘以边缘概率密度 $p(\mathbf{w})$——概率密度函数的乘法定律

于是，对数似然函数:

$$l(\mathbf{w}) = \log \frac{1}{\sqrt{2\pi\lambda^{-1}}} - \frac{\lambda}{2}\mathbf{w}^t\mathbf{w} + n\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^t\mathbf{x}_i)^2 \tag{7}$$

因此，

$$\max_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^t\mathbf{x}_i)^2 + \frac{\lambda}{2}\mathbf{w}^t\mathbf{w}$$

上式就是带有 L_2 正则项的线性回归。现在请思考，LASSO 问题的概率解释是什么？

2. 正则线性回归的概率解释

观察上述问题中 L_2 范数回归算子的引入。我们假设 $\mathbf{w}$ 满足正太分布，即 $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$。于是有

$$p(\mathbf{w}) = \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(\frac{-\mathbf{w}^t\mathbf{w}}{2\lambda^{-1}}\right)$$

如果 $\mathbf{w}$ 满足其他分布？比如 $\mathbf{w} \sim La(\mathbf{0}, \lambda^{-1}\mathbf{I})$（拉普拉斯分布，Laplace distribution），即

$$p(\mathbf{w}) = \frac{1}{2\lambda^{-1}} \exp\left(-\frac{|\mathbf{w}|}{\lambda^{-1}}\right)$$

于是，可得：

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\mathbf{w}) &= p(\mathbf{y}|\mathbf{X};\mathbf{w})p(\mathbf{w}) \\ &= \frac{1}{2\lambda^{-1}}\exp\left(-\frac{|\mathbf{w}|}{\lambda^{-1}}\right)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(y_i-\mathbf{w}^t\mathbf{x}_i)^2}{2\sigma^2}\right) \end{aligned} \tag{8}$$

同样最大化对数似然函数，可得

$$\max_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^t\mathbf{x}_i)^2 + \frac{\lambda}{2}|\mathbf{w}|$$

由此可知，对于正规线性回归模型来说，其不同的正则项，实际上对应了对 $\mathbf{w}$ 分布的不同假设。

3. 对数几率回归的概率解释

在上一节课内容中，我们说对数几率回归具有如下形式，即

$$y = \frac{1}{1 + e^{-\mathbf{w}^t\mathbf{x}}}$$

然后 $\ln \frac{y}{1-y} = \mathbf{w}^t\mathbf{x}$，我们说 y 可以视为样本 $\mathbf{x}$ 为正例的可能性，而 $1 - y$ 视为样本 $\mathbf{x}$ 为反例的可能性。

3. 对数几率回归的概率解释

从概率角度来说，那么可以有 $y = P(y = 1|\mathbf{x})$，$1 - y = P(y = 0|\mathbf{x})$。于是，

$$\begin{aligned} P(y = 1|\mathbf{x}) &= \frac{1}{1+e^{-\mathbf{w}^t\mathbf{x}}} = \frac{e^{\mathbf{w}^t\mathbf{x}}}{1+e^{\mathbf{w}^t\mathbf{x}}} \doteq P_1(\mathbf{x};\mathbf{w}) \\ P(y = 0|\mathbf{x}) &= 1 - \frac{1}{1+e^{-\mathbf{w}^t\mathbf{x}}} = \frac{1}{1+e^{\mathbf{w}^t\mathbf{x}}} \doteq P_0(\mathbf{x};\mathbf{w}) \end{aligned} \tag{9}$$

于是对于任一样本 $\mathbf{x}_i$ 来说，其属于正例或者反例的概率可以写成：

$$P(y_i|\mathbf{x}_i;\mathbf{w}) = P_1(\mathbf{x}_i;\mathbf{w})^{y_i} P_0(\mathbf{x}_i;\mathbf{w})^{1-y_i} \tag{10}$$

于是，同样定义联合概率密度（似然函数）：

$$P(\mathbf{y}|\mathbf{X};\mathbf{w}) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i;\mathbf{w})$$

最大化对数似然，即

$$\begin{aligned}
\max l(\mathbf{w}) &= \max \textstyle\sum_{i=1}^{n} \ln P(y_i|\mathbf{x}_i;\mathbf{w}) \\
&= \max \textstyle\sum_{i=1}^{n} y_i \ln(P_1(\mathbf{x}_i;\mathbf{w})) + (1-y_i)\ln(P_0(\mathbf{x}_i;\mathbf{w})) \\
&= \max \textstyle\sum_{i=1}^{n} y_i \ln(\frac{e^{\mathbf{w}^t\mathbf{x}_i}}{1+e^{\mathbf{w}^t\mathbf{x}_i}}) + (1-y_i)\ln(\frac{1}{1+e^{\mathbf{w}^t\mathbf{x}_i}}) \\
&= \max \textstyle\sum_{i=1}^{n} \left(\mathbf{w}^t\mathbf{x}_i y_i - \ln(1+e^{\mathbf{w}^t\mathbf{x}_i})\right) \\
&= \min \textstyle\sum_{i=1}^{n} \left(-\mathbf{w}^t\mathbf{x}_i y_i + \ln(1+e^{\mathbf{w}^t\mathbf{x}_i})\right) = \min \tilde{l}(\mathbf{w})
\end{aligned} \tag{11}$$

如何求解上式（11）——梯度下降法。

牛顿下降法（Newton method）

思想：

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \tag{12}$$

假设 $\Delta x \to 0$，则可以得到 $f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 = 0$，于是 $\Delta x = -2\frac{f'(x)}{f''(x)}$。于是可以得到更新规则：

$$x \doteq x - \frac{f'(x)}{f''(x)} \tag{13}$$

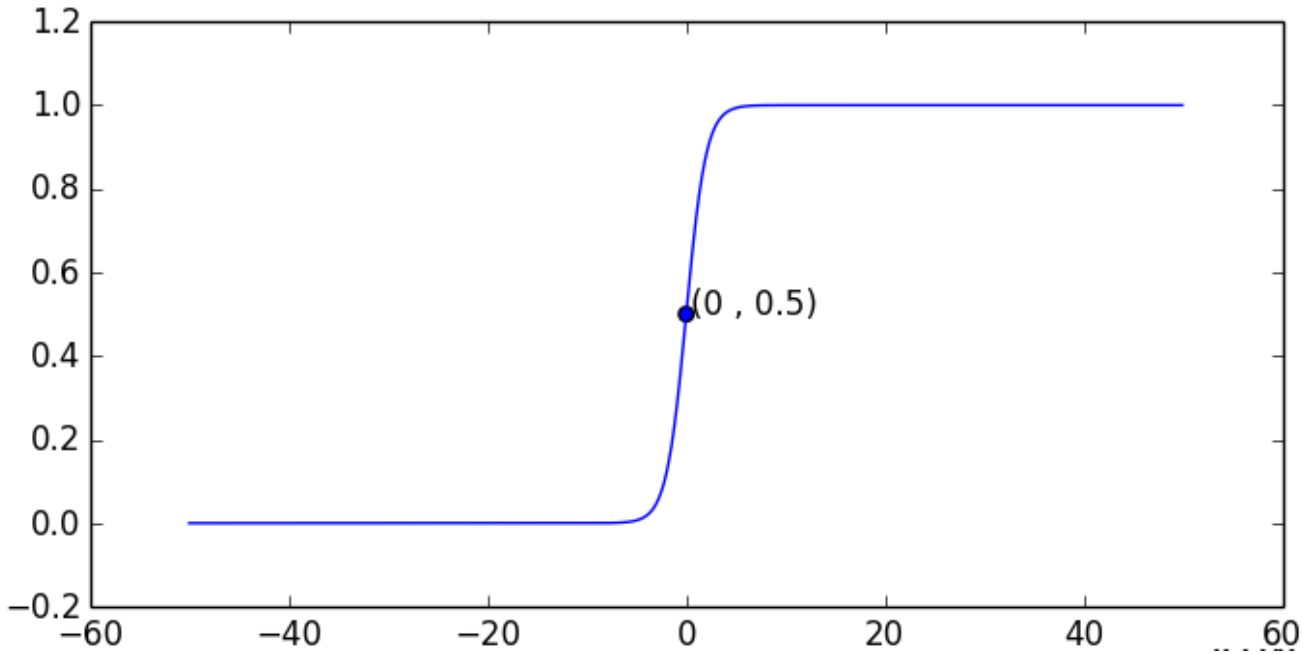
于是对数几率回归参数 $\mathbf{w}$ 的更新规则，可以表示为：

$$\mathbf{w} \doteq \mathbf{w} - \left(\frac{\partial \tilde{l}(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^t}\right)^{-1} \frac{\partial \tilde{l}(\mathbf{w})}{\partial \mathbf{w}} \tag{14}$$

请尝试计算 $\frac{\partial \tilde{l}(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^t}$ 以及 $\frac{\partial \tilde{l}(\mathbf{w})}{\partial \mathbf{w}}$(exercise!!)。

4. 线性判别分析 (Linear discriminant analysis, LDA)

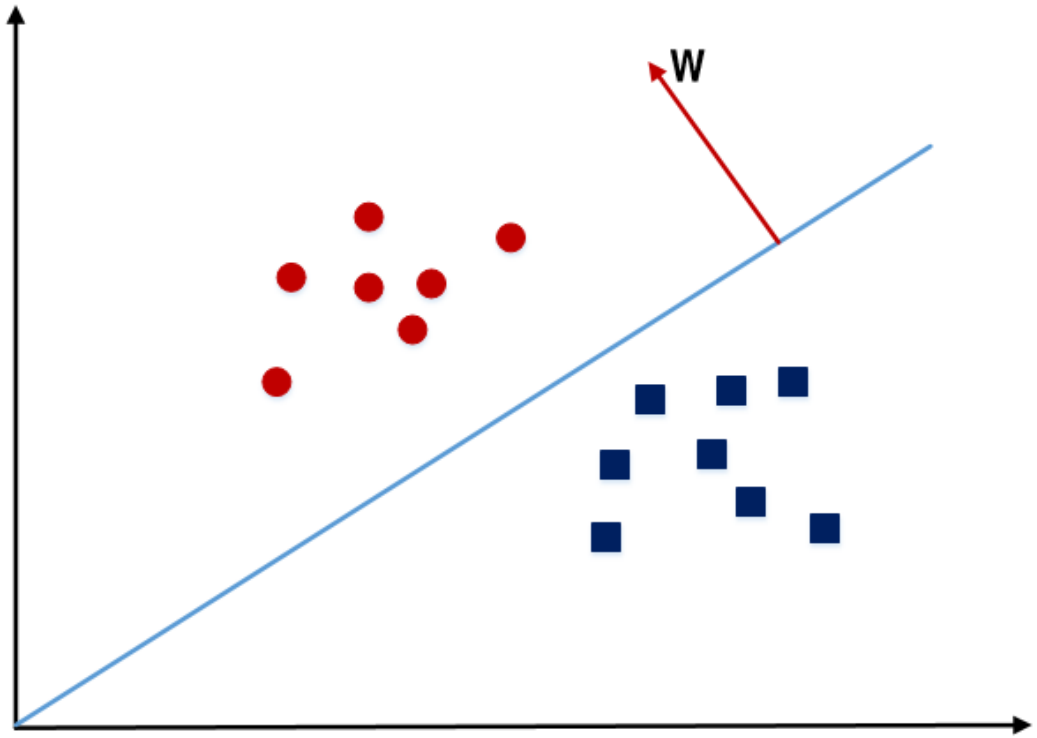
回顾对数几率函数 $y = \frac{1}{1+e^{-z}}$ 如下图，



可见，当 $z > 0$ 时，$y \to 1$。回顾对数几率回归 $y = \frac{1}{1+e^{-\mathbf{w}^t\mathbf{x}}}$，那么 $\mathbf{w}^t\mathbf{x} > 0$，$y \to 1$。

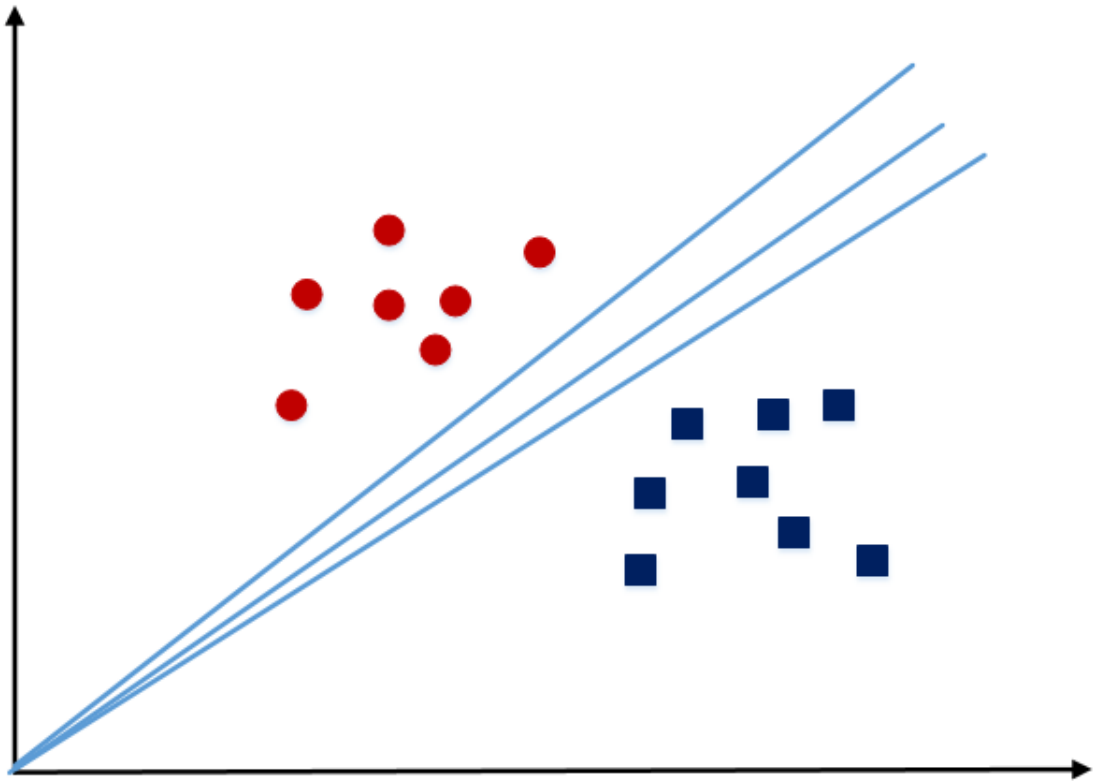
4. 线性判别分析 (Linear discriminant analysis, LDA)

我们知道，$\mathbf{w}^t\mathbf{x} = 0$ 表示一条直线，我们观察二维平面中直线形态：$Ax + By = 0 \Leftrightarrow \mathbf{w}^t\mathbf{x} = 0$，这里 $\mathbf{x} = (x, y)^t$，$\mathbf{w} = (A, B)^t$。



4. 线性判别分析 (Linear discriminant analysis, LDA)

现在，来考虑如何找到一条最好的直线能够将两组样本分开。

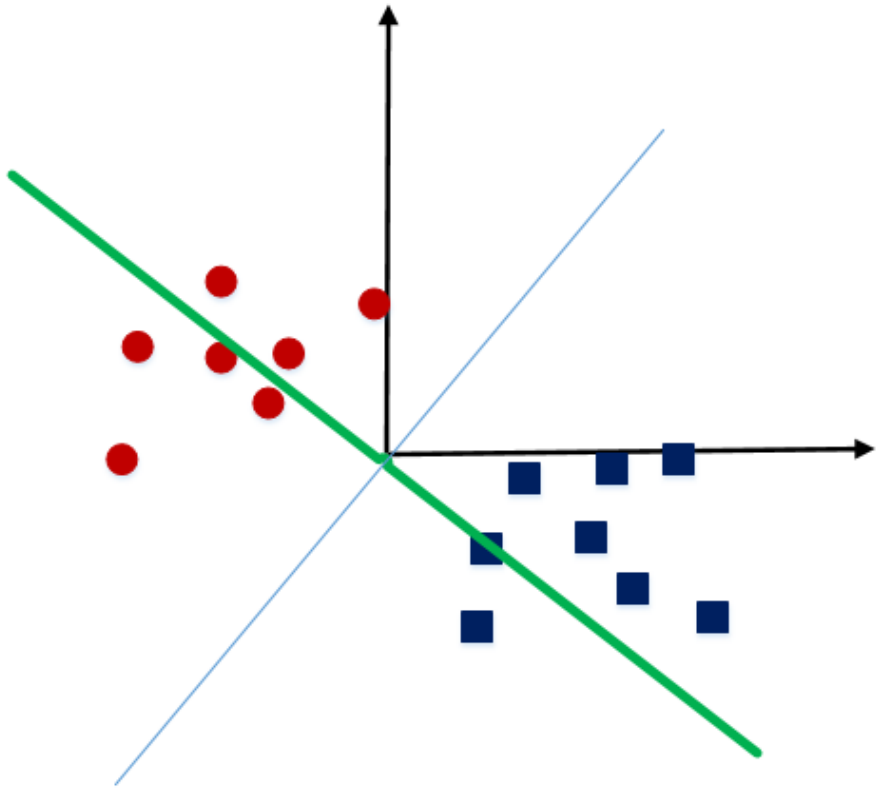


4. 线性判别分析 (Linear discriminant analysis, LDA)

考虑一个向量投影问题。以二维平面为例，如果直线为 $Ax + By = 0$，则法向量为 $(A, B)^t$，方向向量为 $(-B, A)^t$。则任一向量 $\mathbf{x} = (x, y)^t$ 在直线 $Ax + By = 0$ 上的投影等于其方向向量点乘上该向量，即为 $-Bx + Ay$ (exercise!!)。由此推广，**对于高维空间中，任一向量 $\mathbf{w}$，该空间中任一样本 $\mathbf{x}$ 在该向量方向上的投影等于 $\mathbf{w}^t\mathbf{x}$。**

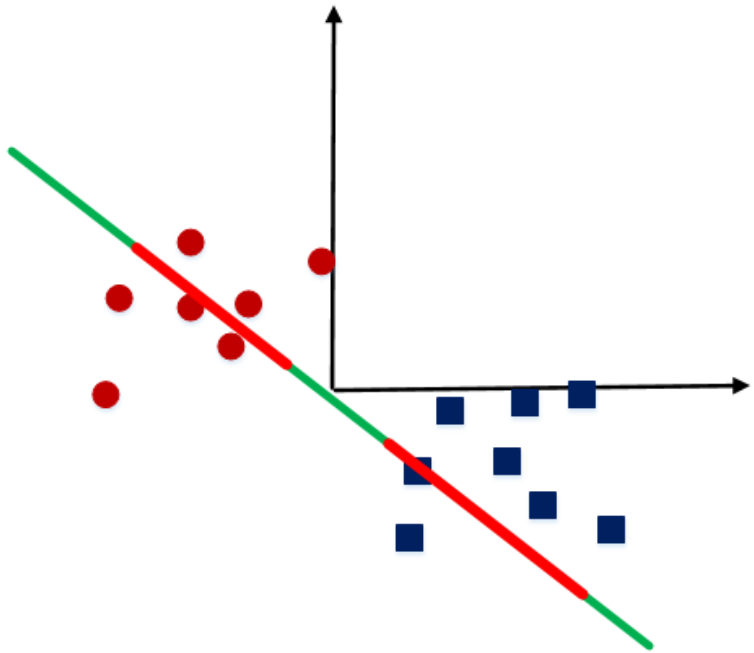
4. 线性判别分析 (Linear discriminant analysis, LDA)

观察如下图，假设其中蓝色直线为最好的分割线，满足 $Ax + By = 0$，绿色直线为与其正交的直线，则该直线方程可以表示为 $Bx - Ay = 0$，（法向量为 $(B, -A)$，方向向量为 (A, B)）。则通过计算出绿色直线，就可以计算出蓝色直线。



4. 线性判别分析 (Linear discriminant analysis, LDA)

现在假设绿色直线方向向量为 $\mathbf{w}$，则任一样本 $\mathbf{x}$ 投影点为 $\mathbf{w}^t\mathbf{x}$。下图中红色圆点为一类，蓝色方块为令一类。可以计算两类在绿色直线上的投影区域。一个最好的结果是，这两个投影区域分开的越远越好，同时为了保证同一类样本的紧致性，同类样本应该聚的越紧越好。



于是，为了满足上述目的，考虑目标函数如下

$$\begin{aligned} J &= \frac{\|\mathbf{w}^t\mu_0 - \mathbf{w}^t\mu_1\|_2^2}{\mathbf{w}^t\mathbf{M}_0\mathbf{w} + \mathbf{w}^t\mathbf{M}_1\mathbf{w}} \\ &= \frac{\mathbf{w}^t(\mu_0-\mu_1)(\mu_0-\mu_1)^t\mathbf{w}}{\mathbf{w}^t(\mathbf{M}_0+\mathbf{M}_1)\mathbf{w}} \end{aligned} \tag{15}$$

其中，μ_0, μ_1 分别为两类的均值，$\mathbf{M}_0, \mathbf{M}_1$ 分别为两类的协方差（矩阵）。定义类间离散度矩阵 $\mathbf{S}_b$ 以及类内离散度矩阵 $\mathbf{S}_w$:

$$\begin{aligned} \mathbf{S}_b &= (\mu_0 - \mu_1)(\mu_0 - \mu_1)^t \\ \mathbf{S}_w &= \sum_{\mathbf{x}\in\mathbf{X}_0}(\mathbf{x} - \mu_0)(\mathbf{x} - \mu_0)^t + \sum_{\mathbf{x}\in\mathbf{X}_1}(\mathbf{x} - \mu_1)(\mathbf{x} - \mu_1)^t \end{aligned} \tag{16}$$

4. 线性判别分析 (Linear discriminant analysis, LDA)

LDA 目标函数可以表示成：

$$\max_{\mathbf{w}} J = \max_{\mathbf{w}} \frac{\mathbf{w}^t \mathbf{S}_b \mathbf{w}}{\mathbf{w}^t \mathbf{S}_w \mathbf{w}} \tag{17}$$

上式即 $\mathbf{S}_b$ 与 $\mathbf{S}_w$ 的广义瑞利商 (generalized Rayleigh quotient)。

上式又可以转换成：

$$\begin{aligned} &\max_{\mathbf{w}} && \mathbf{w}^t \mathbf{S}_b \mathbf{w} \\ &s.t. && \mathbf{w}^t \mathbf{S}_b \mathbf{w} = 1 \end{aligned} \tag{18}$$

通过拉格朗日乘子法，上式又等价于

$$\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w} \tag{19}$$

其中 λ 为拉格朗日乘子。由此可知，$\mathbf{w}$ 为广义矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 最大特征值所对应的特征向量。

4. 线性判别分析 (Linear discriminant analysis, LDA)

最后说明，LDA 方法不仅仅适用于二分类问题，也可以用于多分类问题。用于多分类问题是，只需对 $\mathbf{S}_w, \mathbf{S}_b$ 做简单的修改即可。

Thanks !