

1. 有监督学习

- 1.1 介绍
- 1.2 符号和概念
- 1.3 线性模型(Linear models)
 - 1.3.1 线性回归(Linear regression)
 - 1.3.2 分类和逻辑回归
 - 1.3.3 广义线性模型(Generalized Linear Models)
- 1.4 支持向量机(Support Vector Machines)
- 1.5 生成学习(Generative Learning)
 - 1.5.1 高斯判别分析(Gaussian Discriminant Analysis)
 - 1.5.2 朴素贝叶斯(Naive Bayes)
- 1.6 基于树方法和集成方法
- 1.7 其他非参数方法
- 1.8 学习理论

1. 有监督学习

1.1 介绍

给定数据点集合 $\{x^{(1)}, \dots, x^{(m)}\}$ 和输出集合 $\{y^{(1)}, \dots, y^{(m)}\}$ ，我们想建立一个分类器，让它学习如何从 x 预测 y 。

- 预测类型(Type of prediction)

不同类型的预测模型见下表：

模型	输出	例子
回归	连续	线性回归
分类器	类别	逻辑回归、支持向量机、朴素贝叶斯

- 模型类型(Type of model)

模型	目标	例子
判别式	估计 $P(y x)$	各种回归、支持向量机
生成式	预测 $P(x y)$ 用于推断 $P(y x)$	GDA，朴素贝叶斯

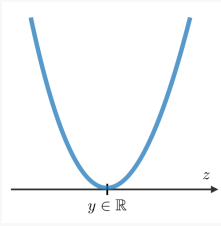
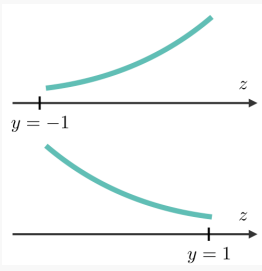
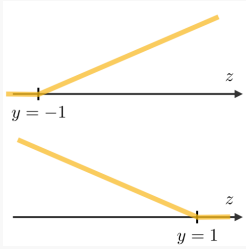
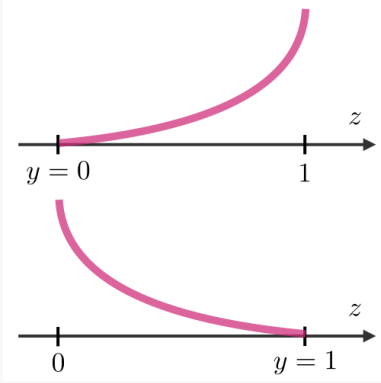
1.2 符号和概念

- 假设(Hypothesis)

记一个假设 h_θ ，是我们选择的模型。对于给定的输入 $x^{(i)}$ ，模型预测结果是 $h_\theta(x^{(i)})$ 。

- 损失函数(Loss function)

损失函数 $L : (z, y) \in \mathbb{R} \times Y \mapsto L(z, y) \in \mathbb{R}$ ，把预测值 y 和真实值 z 作为输入，输出他们的差异程度。常见的损失函数见下表：

二乘法	Logistic	Hinge	交叉熵
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-[y \log(z) + (1 - y) \log(1 - z)]$
			
线性回归	逻辑回归	支持向量机	神经网络

- 代价函数(Cost function)

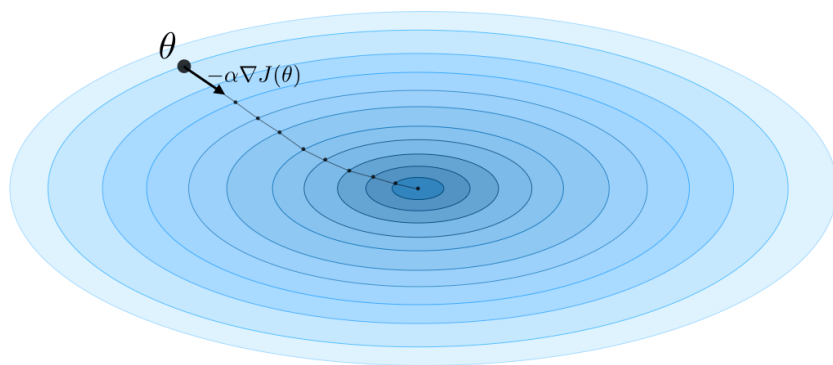
代价函数 J 通常用于表示模型的性能，和损失函数 L 一起，定义如下：

$$J_{\theta} = \sum_{i=1}^m L(h_{\theta}(x^{(i)}, y^{(i)}))$$

- 梯度下降法(Gradient descent)

若学习率 $\alpha \in \mathbb{R}$ ，梯度下降法更新规则用学习率和代价函数 J 表示：

$$\theta \leftarrow \theta - \alpha \nabla J(\theta)$$



备注：随机梯度下降算法(Stochastic gradient descent, SGD)基于每个训练数据更新参数，而批量梯度下降法(batch gradient descent)是基于批量数据。

- 似然(Likelihood)

模型的似然 $L(\theta)$ 是通过将其最大化找到最优的参数 θ 。在实际过程中，我们一般用对数似然 $\ell(\theta) = \log(L(\theta))$ ，更容易优化，表示如下：

$$\theta^{opt} = \arg \max_{\theta} L(\theta)$$

- 牛顿迭代法(Newton's algorithm)

牛顿迭代法是一种数值方法，找到一个 θ ，使 $\ell'(\theta) = 0$ 。它的更新规则如下：

$$\theta \leftarrow \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$

备注：多维泛化(multidimensional generalization)，也被称作牛顿-拉夫逊迭代法(Newton-Raphson method)，更新规则如下：

$$\theta \leftarrow \theta - (\nabla_{\theta}^2 \ell(\theta))^{-1} \nabla_{\theta} \ell(\theta)$$

1.3 线性模型(Linear models)

1.3.1 线性回归(Linear regression)

我们假设 $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$ 。

- 正则方程(Normal equations)

矩阵 X , 代价函数最小值 θ 是一个闭式方案：

$$\theta = (X^T X)^{-1} X^T y$$

- 最小二乘法(LMS algorithm)

记学习率 α ，对于 m 个数据点的训练集，最小二乘法更新规则（也叫做Widrow-Hoff学习规则），如下：

$$\forall, \theta_j \leftarrow +\alpha \sum_{i=1}^m [y^{(i)} - h_{\theta}(x^{(i)})] x_j^{(i)}$$

- 局部加权回归(Locally Weighted Regression)

局部加权回归，即LWR，是线性回归的一种变体，它将每个训练样本的代价函数加权为 $\omega^{(i)}(x)$ ，用参数 $\tau \in \mathbb{R}$ 可定义为：

$$\omega^{(i)}(x) = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

1.3.2 分类和逻辑回归

- 激活函数(Sigmoid function)

激活函数 g ，即逻辑函数，定义如下：

$$\forall z \in \mathbb{R}, g(z) = \frac{1}{1 + e^{-z}} \in [0, 1]$$

- 逻辑回归(Logistic regression)

假设 $y|x; \theta \in \text{Bernoulli}(\phi)$ ，表示如下：

$$\phi = p(y = 1|x; \theta) = \frac{1}{1 + \exp(-\theta^T x)} = g(\theta^T x)$$

备注：逻辑回归中没有闭式方案。

- SOFTMax回归(Softmax regression)

SOFTMax回归，也被称为多元逻辑回归，用于处理输出类别多于2个的多分类问题。按照惯例，设 $\theta_K = 0$ ，每个类 i 的伯努力参数 ϕ_i ：

$$\phi_i = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^K \exp(\theta_j^T x)}$$

1.3.3 广义线性模型(Generalized Linear Models)

- 指数族(Exponential family)

如果一个分布可以用自然参数表示，那么这类分布可以叫做指数族，也被称为正则参数(canonical parameter)或连接函数(link function)。记 η ，素数统计量 $T(\eta)$ 和对数划分函数 $\alpha(\eta)$ ，表示如下：

$$p(y; \eta) = b(y) \exp(\eta T(y) - \alpha(\eta))$$

备注：通常情况 $T(y) = y$ 。同样的， $\exp(-\alpha(\eta))$ 可以看作正则化参数，使得概率结果是1。

常用的指数分布见下表：

分布	η	$T(y)$	$\alpha(\eta)$	$b(\eta)$
伯努力	$\log(\frac{\phi}{1-\phi})$	y	$\log(1 + \exp(\eta))$	1
高斯	μ	y	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})$
泊松	$\log(\lambda)$	y	e^η	$\frac{1}{y!}$
几何	$\log(1 - \phi)$	y	$\log(\frac{e^\eta}{1-e^\eta})$	1

- 广义线性模型(Assumptions of GLMs)

广义线性模型的目标是预测一个随机变量 y ，作为 $x \in \mathbb{R}$ 的函数，并且依赖于下面3个假设

$$(1) \boxed{y|x; \theta \sim \text{ExpFamily}(\eta)} \quad (2) \boxed{h_\theta(x) = E[y|x; \theta]} \quad (3) \boxed{\eta = \theta^T x}$$

备注：普通最小二乘法和逻辑回归是GLM的特例。

1.4 支持向量机(Support Vector Machines)

支持向量机是为了找到一条线，使最小距离最大化。

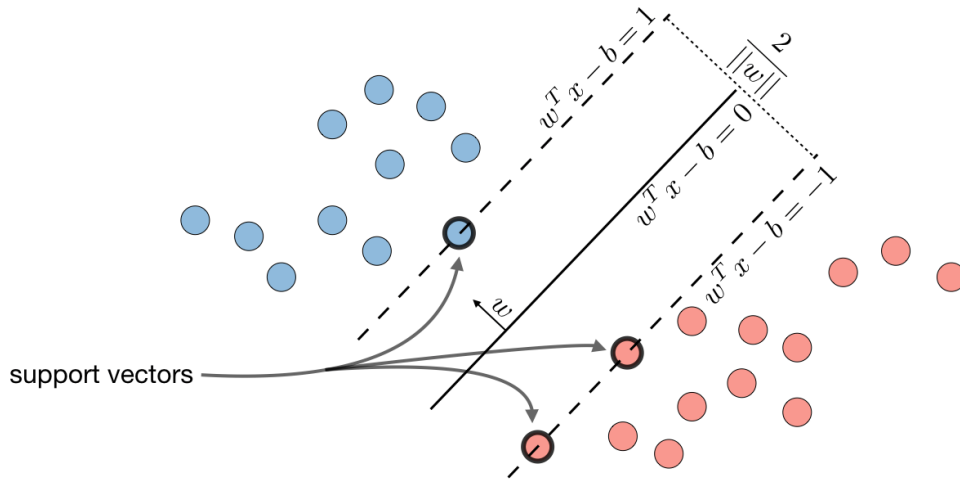
- 最优间隔分类器(Optimal margin classifier)

最优间隔分类器 h 定义如下：

$$h(x) = \text{sign}(\omega^T x - b)$$

其中 $(\omega, b) \in \mathbb{R}^n \times \mathbb{R}$ 是下面两个最优问题的解：

$$\min \frac{1}{2} \|\omega\|^2 \quad \text{使得} \quad y^{(i)} (\omega^T x^{(i)} - b) \geq 1$$



备注：线定义 $\omega^T x - b = 0$ 。

- Hinge损失

支持向量机的Hinge损失定义如下：

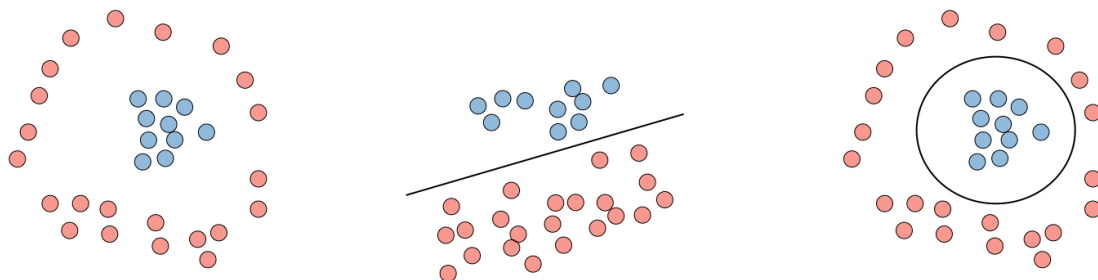
$$L(z, y) = [1 - yz]_+ = \max(0, 1 - yz)$$

- 核(Kernel)

给定特征映射 ϕ ，核 K 定义如下：

$$K(x, z) = \phi(x^T) \phi(z)$$

在实际问题中，高斯核 $K(x, z) = \exp(-\frac{\|x-z\|^2}{2\sigma^2})$ 最常用：



Non-linear separability \longrightarrow Use of a kernel mapping ϕ \longrightarrow Decision boundary in the original space

备注：我们使用“核技巧”去计算损失函数，因为我们不需要知道明确的图 ϕ ，通常非常复杂。相反的，只需要 $K(x, z)$ 的值。

- 拉格朗日(Lagrangian)

我们定义拉格朗日 $\mathcal{L}(\omega, b)$ 如下：

$$\mathcal{L}(\omega, b) = f(\omega) + \sum_{i=1}^l \beta_i h^i(\omega)$$

备注：系数 β_i 称为拉格朗日乘数。

1.5 生成学习(Generative Learning)

生成模型首先尝试通过估计 $P(x|y)$ 去了解数据如何生成，而后我们可以通过贝叶斯规则估计 $P(y|x)$ 。

1.5.1 高斯判别分析(Gaussian Discriminant Analysis)

- 设置

高斯判别分析假设存在 y 并且 $x|y=0$ 、 $x|y=1$ ，满足：

$$y \sim \text{Bernoulli}(\phi)$$

$$x|y=0 \sim \mathcal{N}(\mu_0, \Sigma)$$

$$x|y=1 \sim \mathcal{N}(\mu_1, \Sigma)$$

- 估计

最大似然估计统计如下：

$\hat{\phi}$	$\hat{\mu}_j (j=0,1)$	$\hat{\Sigma}$
$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=1\}}$	$\frac{\sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=j\}}}$	$\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$

1.5.2 朴素贝叶斯(Naive Bayes)

- 假设

朴素贝叶斯假设每个数据点的特征都相互独立的：

$$P(x|y) = P(x_1, x_2, \dots | y) = P(x_1|y)P(x_2|y) \dots = \prod_{i=1}^n P(x_i|y)$$

- 求解

当 $k \in \{0, 1\}, l \in [1, L]$ 时，最大似然估计给出了如下解决方案：

$$P(y=k) = \frac{1}{m} \times \#\{j|y^{(j)}=k\}$$

$$P(x_i=l|y=k) = \frac{\#\{j|y^{(j)}=k \text{ and } x_i^{(j)}=l\}}{\#\{j|y^{(j)}=k\}}$$

备注：朴素贝叶斯广泛用于文字分类。

1.6 基于树方法和集成方法

即可用于回归，又可用于分类的方法。

- 决策树

分类和回归树 (CART), 非常具有可解释性特征。

- Boosting

其思想就是结合多个弱学习器，形成一个较强的学习器。

- 随机森林

在样本和所使用的特征上采用Bootstrap，与决策树不同的是，其可解释性较弱。

1.7 其他非参数方法

- k-近邻法(k-nearest neighbors)

数据点的特性由它周围 k 个邻居决定。

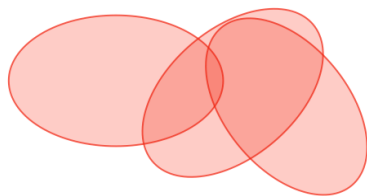
备注：参数 k 越高，偏差越大；参数 k 越低，变量越高。

1.8 学习理论

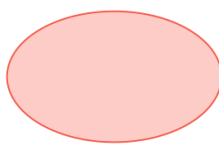
- 联合界(Union bound)

假设 A_1, \dots, A_k 是 k 个事件，则有：

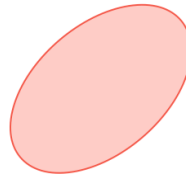
$$P(A_1 \cup \dots \cup A_k) \leq P(A_1) + \dots + P(A_k)$$



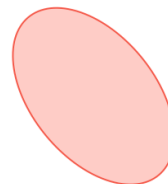
$A_1 \cup A_2 \cup A_3$



A_1



A_2



A_3

- Hoeffding不等式(Hoeffding inequality)

假设 Z_1, \dots, Z_m 是伯努力分布 ϕ 的 m 个变量。假设 $\hat{\phi}$ 是他们采样均值，并且固定 $\gamma > 0$ 。我们得到：

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp(-2\gamma^2 m)$$

备注：此不等式又叫做切尔诺绑定(Chernoff bound)。

- 训练误差(Training error)

对于给定的分类器 h ，我们定义训练误差为 $\hat{\epsilon}(h)$ ，也被称作“经验风险”或“经验误差”，如下所示：

$$\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^m 1_{\{h(x^{(i)}) \neq y^{(i)}\}}$$

- Probably Approximately Correct (PAC)

PAC是经过无数结果在学习理论得到验证的框架，有如下假设：

- 测试和训练集合遵循相同的分布
- 训练样本是独立绘制的

- 样本打散(Shattering)

给定集合 $S = \{x^{(1)}, \dots, x^{(d)}\}$ ，给定分类器 \mathcal{H} ，我们可以说 \mathcal{H} 用标签 $\{y^{(1)}, \dots, y^{(d)}\}$ 打散 S ，我们得到：

$$\boxed{\exists h \in \mathcal{H}, \forall i \in [1, d], h(x^{(i)}) = y^{(i)}}$$

- 上限法(Upper bound theorem)

假设 \mathcal{H} 是一个奈特假说类， $|\mathcal{H}| = k$ ，假设 δ 和样本数量 m 是固定的。然后，概率至少 $1 - \delta$ ，我们得到：

$$\boxed{\epsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \epsilon(h) \right) + 2\sqrt{\frac{1}{2m} \log\left(\frac{2k}{\delta}\right)}}$$

- VC维(VC dimension)

给定假设类 \mathcal{H} 的VC(Vapnik-Chervonenkis)维，用 $VC(\mathcal{H})$ 表示 \mathcal{H} 打散的最大集合：

备注：如果 \mathcal{H} 是2维空间的线性分类器集合，则 \mathcal{H} 的VC维是3。



- Theorem (Vapnik)

假设给定的 \mathcal{H} ，其中 $VC(\mathcal{H}) = d$ ，训练样本数量 m 。最小概率 $1 - \delta$ ，我们得到：

$$\boxed{\epsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \epsilon(h) \right) + O\sqrt{\frac{d}{m} \log\left(\frac{m}{d}\right) + \frac{1}{m} \log\left(\frac{1}{\delta}\right)}}$$