第一周

监督学习：回归，分类

无监督：聚类

单变量线性回归：

~~方法~~ $h_\theta(x) = \theta_0 + \theta_1 x$

Loss：$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Hypothesis：$h_\theta(x) = \theta_0 + \theta_1 x$

Pamameters：θ_0, θ_1

Cost function：$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Goal：$\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

梯度下降：

batch gradient descent：

repeat until convergence{

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ $j = 0$ or 1

}

注：需同时更新 θ_0, θ_1

多变量线性回归：

Hypothesis：$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

or $h_\theta(x) = \theta^T X$ 其中 $x_0 = 1$

Loss：$J(\theta_0, \theta_1, \theta_2 \cdots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Parameters: $\theta_0, \theta_1, \cdots \theta_n$

Cost Function: $J(\theta_0, \theta_1, \cdots \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

Goal: $\underset{\theta_0, \theta_1, \cdots \theta_n}{\text{minimize}} \; J(\theta_0, \theta_1, \cdots \theta_n)$

梯度下降法tips:

① Feature scaling: $x_n = \frac{x_n - \mu_n}{s_n}$ 其中 μ_n 为平均值, s_n 为标准差.

② Learning rate: 如何调整

特征与多项式回归:

可以添加多项式项整合特征联系

注: 采用多项式回归, 特征缩放非常有必要。

Normal Equation (正规方程):

用于求解最优解: $\theta = (X^T X)^{-1} X^T y$

注: 仅适用于可逆矩阵

梯度下降	VS	正规方程
需要选择 α		不需要
多次迭代		一次运算
适于 n 大		需计算 $(X^T X)^{-1}$, $O(n^3)$, $n<10000$ 可接受
各种模型		只适用于线性模型.

逻辑回归：(Logistic Regression)

① 分类问题

△ 为什么是线性分类？

① $\theta \cdot x$ 为 x 的线性组合

② 决策边界是线性的.

$\theta \cdot x = 0$

Hypothesis：$h_\theta(x) = g(\theta^T x)$

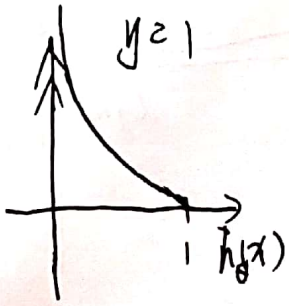
$$g(z) = \frac{1}{1+e^{-z}} \quad (\text{Sigmoid Function})$$

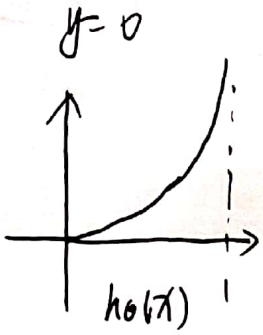
决策边界(decision boundary)

Loss：如果用平方误差将 $g(z)$ 代入，得到一个非凸函数

$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)})$，其中：

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = \begin{cases} -\log(h_\theta(x)), & \text{if } y=1 \\ -\log(1-h_\theta(x)), & \text{if } y=0 \end{cases}$$





缩写为：$Cost(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$

注：Feature scaling is important.

其它求解法：共轭梯度，局部优化法，有限内存局部优化法

多类别分类：训练多个分类器.

正则化：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left[(h_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{n}\theta_j^2\right]$

神经网络：

Loss Function：

$$J(\Theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{k}y_k^{(i)}\log(h_\theta(x^{(i)}))_k+(1-y_k^{(i)})\log(1-(h_\theta(x^{(i)}))_k)\right]+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

反向传播算法：☆

梯度检验：

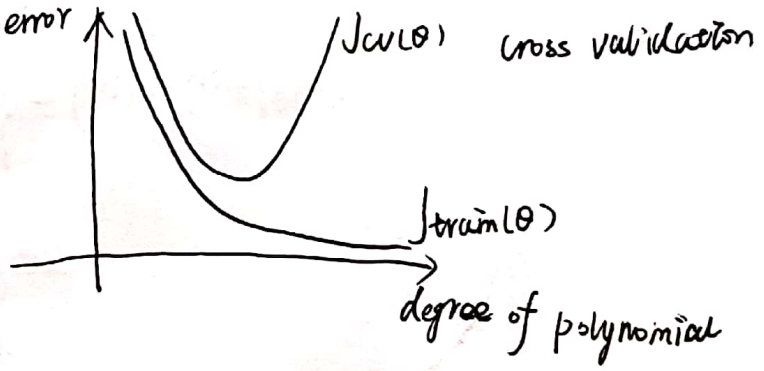
检验 θ 处斜率与反向传播计算差距.

注要随机初始化.

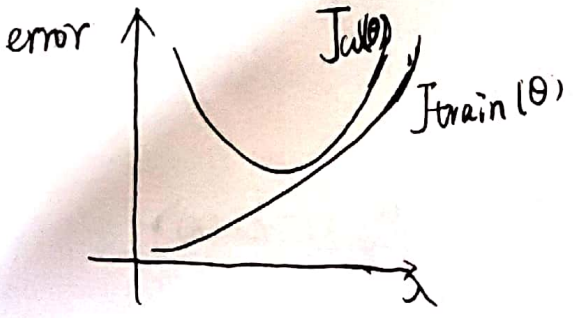
偏差与方差的诊断：

高偏差：欠拟合

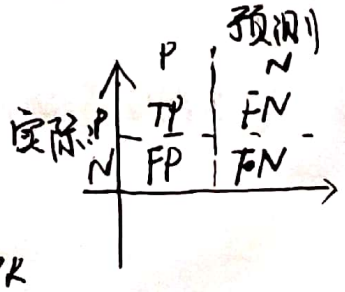
高方差：过拟合.



正则参数的选择：



类偏斜：



准确率 $P=\frac{TP}{TP+FP}$

召回率 $R=\frac{TP}{TP+FN}$

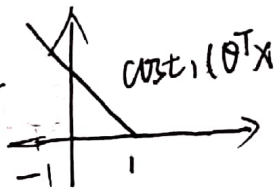
F_1 score：$\frac{2PR}{R+P}$

（监督学习）

Loss function: $CA+B$

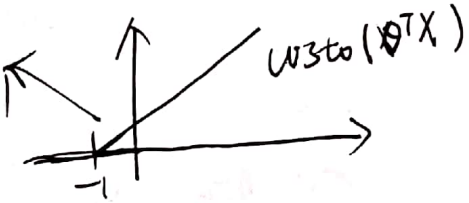
$$\min_{\theta} C\sum_{i=1}^{m}\left[y^{(i)}\text{cost}_1(\theta^T x^{(i)})+(1-y^{(i)})\text{cost}_0(\theta^T x^{(i)})\right]+\frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

Hypothesis: $h_\theta(x)=\begin{cases}1 & \text{if } \theta^T x\geq 0\\ 0 & \text{otherwise}\end{cases}$



（大间距分类器）：if $y=1$, we want $\theta^T x\geq 1$

if $y=0$, we want $\theta^T x\leq -1$



核函数：与标记点

线性核；高斯核、多项式核、字符串核、卡方核.

Logestic regression vs ~~SVM~~ SVM

n: number of feature m number of examples

① n large (relative to m) Logestic or SVM without kernel.

② n is small, m intermediate. SVM with Gaussian kernel

③ n is small m is large, Create more feature, then use Logestic or SVM without a kernel.

K-均值：① 随机选取K个聚类中心

② 代分类别

③ 更新聚类中心

④ 直至中心不变。

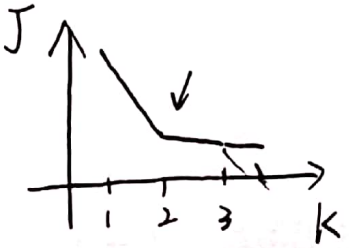
Loss Function: $J(c^{(1)}, c^{(2)}, \cdots c^{(m)}, u_1 \cdots u_K)$

$$= \frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - u_{c^{(i)}} \|^2$$

$u_{c^{(i)}}$ 代表与 $x^{(i)}$ 最近的聚类中心。

选择K值：

使用"肘部法则"



距离/相似度度量：

① 闵可夫斯基距离 $dist(X,Y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p \right)^{\frac{1}{p}}$

② 杰卡德相似系数：

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

③ 余弦相似度：

$$\cos(\theta) = \frac{x^T y}{\|x\| \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

④ Pearson 相似系数：

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-u_X)(Y-u_Y)]}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} (x-u_X)(y-u_Y)}{\sqrt{\sum_{i=1}^{n} (x-u_X)^2} \sqrt{\sum_{i=1}^{n} (y-u_Y)^2}}$$

均一性 p，完整性 r，V-measure：$V=\frac{(1+\beta^2)\times p\times r}{\beta^2*p+r}$

轮廓系数：

降维：

主成分分析（PCA）：

找一个向向量，希望投射平均均方误差尽可能小。

① 均值归一化 $x_j=\frac{x_j-u_j}{\sigma^2}$

② 协方差矩阵 $\Sigma=\frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$

③ 计算 Σ 的特征向量。

$[u,s,v]=svd(\Sigma)$

选取 u 中前 k 个向量，计算新特征向量 $z^{(i)}=u_{reduce}^T * x^{(i)}$

主成分分析 主分数量选择：

$$\rho=\frac{\text{投射的平均均方误差}}{\text{训练集方差}}=\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}-x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2}\leq 1\%$$

则保留 99% 的偏差。

$x_{approx}^{(i)}=u_{reduce}*z^{(i)}$

异常检测：

通过高斯分布来解决异常。

多元高斯：

构建特征的协方差矩阵，用所有特征一起来计算 $p(x)$。

原高斯

不能捕捉特征之间相关性，可以通过组合方法解决

计算成本低。

多元高斯

自动捕捉特征之间相关性。

计算代价大。

必须有 $m>n$，保证协差矩阵不可逆。

推荐系统：

① 基于内容的推荐：已知物品特征。

② 协同过滤：拥有用户参数。亦可同时学习物品和用户参数。

均值归一化解决新用户；预测时：$(\theta^{(j)})^T x^{(i)}+\mu_i$，但面对于新用户评分为均值（大众口味）

随机梯度下降；小批量梯度下降

上限分析：