1 Linear

1.1 Regression

损失函数	$\|\|\mathbf{g}(\mathbf{X})-\mathbf{y}\|\|_2^2$
预测函数	$\mathbf{g}(\mathbf{X})=\mathbf{X}\mathbf{w}+\mathbf{b}=\hat{\mathbf{X}}\hat{\mathbf{w}}$

其中 $\hat{\mathbf{X}}=\{\mathbf{X},1\}$ $\hat{\mathbf{w}}=\{\mathbf{w},b\}$

优化:

$$\min_{\hat{\mathbf{w}}}\frac{1}{2}(\hat{\mathbf{X}}\hat{\mathbf{w}}-\mathbf{y})^\top(\hat{\mathbf{X}}\hat{\mathbf{w}}-\mathbf{y}) \tag{1}$$

偏导:

$$\frac{\partial\cdot}{\partial\hat{\mathbf{w}}}=\hat{\mathbf{X}}^\top(\hat{\mathbf{X}}\hat{\mathbf{w}}-\mathbf{y})=0 \tag{2}$$

结果:

$$\hat{\mathbf{w}}=(\hat{\mathbf{X}}^\top\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top\mathbf{y} \tag{3}$$

1.2 Logical

预测函数	$t=\mathbf{X}\mathbf{w}+\mathbf{b}$
	$\mathbf{g}(\mathbf{X})=\frac{1}{1+\exp^{-t}}$

说明:

对数几率	$\ln(\frac{\mathbf{g}(\mathbf{X})}{1-\mathbf{g}(\mathbf{X})})=\mathbf{X}\mathbf{w}+\mathbf{b}=\hat{\mathbf{X}}\hat{\mathbf{w}}$
概率	$p_i=P(y=1\mid x_i)=g(x_i)$
优化函数的左边	$F(\mathbf{x})=\sum\ln(1+e^{\mathbf{x}_i})$
求导	$\frac{\partial F(\mathbf{x})}{\partial\mathbf{x}}=[\frac{e^{\mathbf{x}_1}}{1+e^{\mathbf{x}_1}},\frac{e^{\mathbf{x}_2}}{1+e^{\mathbf{x}_2}}\cdots\frac{e^{\mathbf{x}_n}}{1+e^{\mathbf{x}_n}}]^\top$ $=[\frac{1}{1+e^{-\mathbf{x}_1}},\frac{1}{1+e^{-\mathbf{x}_2}}\cdots\frac{1}{1+e^{-\mathbf{x}_n}}]^\top$
所以	$\frac{\partial F(\hat{\mathbf{X}}\hat{\mathbf{w}})}{\partial\hat{\mathbf{X}}\hat{\mathbf{w}}}=[\frac{1}{1+e^{-\hat{\mathbf{X}}\hat{\mathbf{w}}_1}},\frac{1}{1+e^{-\hat{\mathbf{X}}\hat{\mathbf{w}}_2}}\cdots\frac{1}{1+e^{-\hat{\mathbf{X}}\hat{\mathbf{w}}_n}}]^\top=\mathbf{g}(\mathbf{X})$

优化 (似然函数):

$$\min_{\mathbf{w}} -\prod g(x_i)^{y_i}(1-g(x_i))^{1-y_i} \tag{4}$$

$$\min_{\mathbf{w}} -\sum y_i \ln p_i + (1-y_i)\ln(1-p_i) \tag{5}$$

$$\min_{\mathbf{w}} -\sum y_i \ln(\frac{p_i}{1-p_i}) + \ln(1-p_i) \tag{6}$$

$$\min_{\mathbf{w}} -\sum y_i \ln(\hat{\mathbf{w}}x_i) - \ln(1+e^{\hat{\mathbf{w}}x_i}) \tag{7}$$

$$\min_{\mathbf{w}} F(\hat{\mathbf{X}}\hat{\mathbf{w}}) - \mathbf{y}^\top\hat{\mathbf{X}}\hat{\mathbf{w}} \tag{8}$$

偏导:

$$\begin{aligned} d(F(\hat{\mathbf{X}}\hat{\mathbf{w}}) - \mathbf{y}^\top\hat{\mathbf{X}}\hat{\mathbf{w}}) &= (\frac{\partial F(\hat{\mathbf{X}}\hat{\mathbf{w}})}{\partial\hat{\mathbf{X}}\hat{\mathbf{w}}})^\top d(\hat{\mathbf{X}}\hat{\mathbf{w}}) - \mathbf{y}^\top\hat{\mathbf{X}}d\hat{\mathbf{w}} \\ &= (\mathbf{g}(\mathbf{X})\hat{\mathbf{X}} - \mathbf{y}^\top\hat{\mathbf{X}})d\hat{\mathbf{w}} \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial\cdot}{\partial\hat{\mathbf{w}}} &= (\mathbf{g}(\mathbf{X})^\top\hat{\mathbf{X}} - \mathbf{y}^\top\hat{\mathbf{X}})^\top \\ &= \hat{\mathbf{X}}^\top(\mathbf{g}(\mathbf{X}) - \mathbf{y}) \end{aligned} \tag{10}$$

梯度下降:

$$\mathbf{w}_{new} = \mathbf{w}_{old} - \alpha\hat{\mathbf{X}}^\top(\mathbf{g}(\mathbf{X}) - \mathbf{y}) \tag{11}$$

2 DecisionTree

通过找到最合适的 point 去把数据分为两个子数据
属性类集 $\mathbf{C}$, 第 i 个属性 $\mathbf{C}_i \in \mathbf{C}$, 第 i 个属性取值为 j, 记作 $\mathbf{C}_i = j$
离散属性先转为多列的 0 和 1 看做多个连续属性

2.1 Classifier

数据 $\mathbf{D}$, 标签集 S 遍历所有的 point

数据集被 point 分为两份 $\mathbf{A}$ 和 $\mathbf{B}$, 得到占比 $P(\mathbf{A} \mid \mathbf{D})$,$P(\mathbf{B} \mid \mathbf{D})$ 得到占比

A	$P_A = [P(S = S_1 \mid \mathbf{A}), P(S = S_2 \mid \mathbf{A}) \ldots P(S = S_n \mid \mathbf{A})]$
B	$P_B = [P(S = S_1 \mid \mathbf{B}), P(S = S_2 \mid \mathbf{B}) \ldots P(S = S_n \mid \mathbf{B})]$

得到信息熵

$$En_{poitn} = P(\mathbf{A} \mid \mathbf{D}) P_A^\top \log_2(P_A) + P(\mathbf{B} \mid \mathbf{D}) P_B^\top \log_2(P_B)$$

找出信息熵最小即使最佳的 point

2.2 Regression

标签值 S 数据集被 point 分为两份 $\mathbf{A}$ 和 $\mathbf{B}$,$S_{\mathbf{A}}$ $S_{\mathbf{B}}$
计算 $S_{\mathbf{A}}$ $S_{\mathbf{B}}$ 平方误差。再相加得 E_{point}
取 E_{point} 最小时的 point