机器学习笔记

Leoeon

2016 年 10 月 25 日

Contents

Chapter 1

广义线性模型GLM

$\vec{x}$ 以$p(\vec{y}|\vec{x})$ 概率映射到$\vec{y}$上。

已知或假设$\vec{y}$服从某种形式已知的分布$p(\vec{y}|\theta_1,\cdots,\theta_K)$。

将$\vec{x}$扩充为$\check{x}$，$\check{x}$中的每一分量为1、$\vec{x}$分量一次项、$\vec{x}$分量高次项等等。

将$\check{x}$投影到一组基$\{\vec{w}_1,\cdots,\vec{w}_K\}$ 上$\{z_1=\vec{w}_1\cdot\check{x},\cdots,z_K=\vec{w}_K\cdot\check{x}\}$上，即提取特征。

则只需找到一组满足$\{[-\infty,+\infty]\rightarrow$ 值域$[\theta_k]\}$的映射$\{\theta_k=f_k(z_k)\}$。

训练样本集$\{(\vec{x}_\varsigma,\vec{y}_\varsigma)\}$，由$L=\prod_\varsigma p(\vec{y}_\varsigma|\vec{\theta}_\varsigma)$，求出$\vec{w}^*=\operatorname{argmax}_{\vec{w}} L$

1.1 指数族分布

为求映射$\{\theta_k=f_k(z_k)\}$，若$p(\vec{y}|\theta_1,\cdots,\theta_K)$ 为指数族分布

$$C(\theta)\cdot H(\vec{y})\cdot e^{\sum_k Q_k(\theta_k)\cdot T_k(\vec{y})}=H(\vec{y})\cdot e^{\sum_k \eta_k\cdot T_k(\vec{y})-b(\eta)}$$

可取某个函数$h_k(z_k)=E_y[T_k(\vec{y})]\left(=\frac{\partial b(\eta)}{\partial\eta_k}\right)$，代入得$f_k$

自然联系函数：若将h_k 形式取为$\frac{\partial b}{\partial\eta_k}$，则可得：$z_k=\eta_k=Q_k(\theta_k)$

1.1.1 多项式分布

1.1.1.1 多项式分布

$$\begin{aligned}
p(y_1,\cdots,y_{K-1}|\theta_1,\cdots,\theta_{K-1}) &= \frac{n!}{\prod_{k=1}^K y_k!}\prod_{k=1}^K\theta_k^{y_k} && (\textstyle\sum_{k=1}^K\theta_k=1,\sum_{k=1}^K y_k=n)\\
&= \frac{n!}{\prod_{k=1}^K y_k!}\cdot e^{\sum_{k=1}^K\ln\theta_k\cdot y_k} && (\textstyle\sum_{k=1}^K\theta_k=1,\sum_{k=1}^K y_k=n)\\
&= \theta_K^n\cdot\frac{n!}{\prod_{k=1}^K y_k!}\cdot e^{\sum_{k=1}^{K-1}\ln\frac{\theta_k}{\theta_K}\cdot y_k} && (\textstyle\sum_{k=1}^K\theta_k=1)
\end{aligned}$$

得：

$$\theta_k=\begin{cases}\frac{e^{z_k}}{1+\sum_{j=1}^{K-1}e^{z_j}} & ,\quad k=1,\cdots,K-1\\ \frac{1}{1+\sum_{j=1}^{K-1}e^{z_j}} & ,\quad k=K\end{cases}$$

或：

$$\theta_k=\frac{e^{e^{z_k}}}{\sum_{j=1}^K e^{e^{z_j}}}\quad,\quad k=1,\cdots,K$$

1.1.1.2 多项式分布

（多项式分布特例：$n=1$）

$$\begin{aligned} p(y_1,\cdots,y_{K-1}|\theta_1,\cdots,\theta_{K-1}) &= \prod_{k=1}^{K}\theta_k^{y_k} && (\textstyle\sum_{k=1}^{K}\theta_k=1,\sum_{k=1}^{K}y_k=1)\\ &= e^{\sum_{k=1}^{K}\ln\theta_k\cdot y_k} && (\textstyle\sum_{k=1}^{K}\theta_k=1,\sum_{k=1}^{K}y_k=1)\\ &= \theta_K\cdot e^{\sum_{k=1}^{K-1}\ln\frac{\theta_k}{\theta_K}\cdot y_k} && (\textstyle\sum_{k=1}^{K}\theta_k=1) \end{aligned}$$

得：

$$\theta_k=\begin{cases}\frac{e^{z_k}}{1+\sum_{j=1}^{K-1}e^{z_j}} &, \quad k=1,\cdots,K-1\\ \frac{1}{1+\sum_{j=1}^{K-1}e^{z_j}} &, \quad k=K\end{cases}$$

或：

$$\theta_k=\frac{e^{e^{z_k}}}{\sum_{j=1}^{K}e^{e^{z_j}}} \quad , \quad k=1,\cdots,K$$

1.1.1.3 二项分布

（多项式分布特例：$K=2$）

$$\begin{aligned} p(y|\theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y}\\ &= (1-\theta)^n\cdot\binom{n}{y}\cdot e^{\ln\frac{\theta}{1-\theta}\cdot y}\\ & \quad (y=0,\cdots,n) \end{aligned}$$

得：

$$\theta=\frac{1}{1+e^{-z}}$$

1.1.1.4 伯努利分布(logistic回归)

（多项式分布特例：$K=2$，$n=1$）

$$\begin{aligned} p(y|\theta) &= \theta^y(1-\theta)^{1-y}\\ &= (1-\theta)\cdot e^{\ln\frac{\theta}{1-\theta}\cdot y}\\ & \quad (y=0,1) \end{aligned}$$

得：

$$\theta=\frac{1}{1+e^{-z}}$$

即：

$$\begin{cases} +\infty & \to & 1\\ 0 & \to & \frac{1}{2}\\ -\infty & \to & 0 \end{cases}$$

则：

$$L=\prod_{\varsigma}\theta_{\varsigma}^{y_{\varsigma}}(1-\theta_{\varsigma})^{1-y_{\varsigma}}$$

为$\text{argmax}_{\vec{w}}L$，令$\frac{\partial}{\partial w_k}\log L=0$，得$\sum_{\varsigma}(y_{\varsigma}-\theta_{\varsigma})x_{\varsigma k}=0$
由此求出$\vec{w}$

1.1.2 正态分布

1.1.2.1 二元正态分布

$$\begin{aligned} p(y|\mu,\sigma) &= \tfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \tfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\mu^2}{2\sigma^2}}\cdot e^{\frac{\mu}{\sigma^2}\cdot y-\frac{1}{2\sigma^2}\cdot y^2} \end{aligned}$$

1.1.2.2 线性最小二乘法

（二元正态分布特例）

$p(y|\mu)$ 服从高斯分布$\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2}}$，映射$z=\mu$

则$\mathrm{argmax}_{\vec{w}}L$ 有：$\mathrm{argmin}_{\vec{w}}||\vec{w}\cdot\check{x}-y||_2$

令$\check{X}=\begin{bmatrix} & \vdots & \\ \cdots & \check{x}_{\varsigma k} & \cdots \\ & \vdots & \end{bmatrix}$，$Y=\begin{bmatrix}\vdots\\ y_\varsigma\\ \vdots\end{bmatrix}$，$\vec{w}=\begin{bmatrix}\vdots\\ w_k\\ \vdots\end{bmatrix}$，则$\check{X}^T\check{X}\vec{w}=\check{X}^TY$

1.1.3 其他例子

泊松分布：

$$\begin{aligned} p(y|\lambda) &= \tfrac{\lambda^y}{y!}e^{-\lambda} \\ &= e^{-\lambda}\cdot\tfrac{1}{y!}\cdot e^{\ln\lambda\cdot y} \\ & (y=0,1,2,\cdots) \end{aligned}$$

几何分布：

$$\begin{aligned} p(y|\theta) &= (1-\theta)^{y-1}\theta \\ &= \tfrac{\theta}{1-\theta}\cdot e^{\ln(1-\theta)\cdot y} \\ & (y=0,1,2,\cdots) \end{aligned}$$

指数分布：

$$\begin{aligned} p(y|\lambda,\mu) &= \lambda e^{-\lambda(y-\mu)} \\ &= \lambda e^{\lambda\mu}\cdot e^{-\lambda\cdot y} \\ & (y>\mu) \end{aligned}$$

幂分布：

$$\begin{aligned} p(y|\theta) &= \theta y^{\theta-1} \\ &= \theta\cdot\tfrac{1}{y}\cdot e^{\theta\cdot\ln y} \\ & (0<y<1) \end{aligned}$$

β分布：

$$\begin{aligned} p(y|a,b) &= \tfrac{1}{\beta(a,b)}y^{a-1}(1-y)^{b-1} \\ &= \tfrac{1}{\beta(a,b)}\cdot\tfrac{1}{y(1-y)}\cdot e^{a\cdot\ln y+b\cdot\ln(1-y)} \\ & (0<y<1) \end{aligned}$$

Γ分布：

$$\begin{aligned} p(y|\alpha,\lambda) &= \tfrac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\lambda y} \\ &= \tfrac{\lambda^\alpha}{\Gamma(\alpha)}\cdot e^{-\lambda\cdot y+(\alpha-1)\cdot\ln y} \\ & (y>0) \end{aligned}$$

1.2 SVM

K=1

训练样本$\{(\vec{x}_\varsigma, y_\varsigma)\}(y_\varsigma = \pm 1)$

分类函数$f(\vec{x}) \overset{def}{=} (\vec{w} \cdot \check{x} + b)$

几何间隔$\gamma_\varsigma \overset{def}{=} (\frac{\vec{w}}{||\vec{w}||} \cdot \check{x}_\varsigma + \frac{b}{||\vec{w}||})y_\varsigma$

最大化训练样本集中最小的几何间隔

$$\max_{\vec{w},b} \min_{\varsigma}(\frac{\vec{w}}{||\vec{w}||} \cdot \check{x}_\varsigma + \frac{b}{||\vec{w}||})y_\varsigma$$

等价于

$$\begin{aligned} &\max_{\vec{w},b} \tilde{\gamma} \\ \text{s.t.} \quad &(\tfrac{\vec{w}}{||\vec{w}||} \cdot \check{x}_\varsigma + \tfrac{b}{||\vec{w}||})y_\varsigma \geq \tilde{\gamma} \end{aligned}$$

等价于

$$\begin{aligned} &\max_{\vec{w},b} \tfrac{1}{||\vec{w}||} \\ \text{s.t.} \quad &(\vec{w} \cdot \check{x}_\varsigma + b)y_\varsigma \geq 1 \end{aligned}$$

等价于$(\min\limits_{\vec{w}} \frac{1}{2}||\vec{w}||^2)$

$$\begin{aligned} &L(\vec{w},\vec{c}) = \tfrac{1}{2}||\vec{w}||^2 - \sum_\varsigma c_\varsigma((\vec{w} \cdot \check{x}_\varsigma + b)y_\varsigma - 1) \\ &\min_{\vec{w},b} \max_{\vec{c}} L(\vec{w},\vec{c}) \\ \text{s.t.} \quad &c_\varsigma \geq 0 \end{aligned}$$

（为取$\max\limits_{\vec{c}}$：当$(\vec{w} \cdot \check{x}_\varsigma + b)y_\varsigma > 1$时，$c_\varsigma = 0$。即仅有支持向量起作用）

等价于

$$\begin{aligned} &L(\vec{w},\vec{c}) = \tfrac{1}{2}||\vec{w}||^2 - \sum_\varsigma c_\varsigma((\vec{w} \cdot \check{x}_\varsigma + b)y_\varsigma - 1) \\ &\max_{\vec{c}} \min_{\vec{w},b} L(\vec{w},\vec{c}) \\ \text{s.t.} \quad &c_\varsigma \geq 0 \end{aligned}$$

等价于(由$\frac{\partial}{\partial w_i}L(\vec{w},\vec{c}) = 0$ 得：$\vec{w} = \sum\limits_\varsigma c_\varsigma y_\varsigma \check{x}_\varsigma$；由$\frac{\partial}{\partial b}L(\vec{w},\vec{c}) = 0$ 得：$0 = \sum\limits_\varsigma c_\varsigma y_\varsigma$)

$$\begin{aligned} &L(\vec{c}) = \sum_\varsigma c_\varsigma - \tfrac{1}{2}\sum_{\varsigma_1}\sum_{\varsigma_2} c_{\varsigma_1}c_{\varsigma_2}y_{\varsigma_1}y_{\varsigma_2}\check{x}_{\varsigma_1} \cdot \check{x}_{\varsigma_2} \\ &\max_{\vec{c}} L(\vec{c}) \\ \text{s.t.} \quad &c_\varsigma \geq 0 \\ &\sum_\varsigma c_\varsigma y_\varsigma = 0 \end{aligned}$$

（$f(\vec{x}) = \sum\limits_\varsigma c_\varsigma y_\varsigma \check{x}_\varsigma \cdot \check{x} + b$）

1.2.0.1 核函数[4]

因$\check{x}$可能因高维导致维度灾难，故可用核函数替代$\check{x}_1 \cdot \check{x}_2$

如：$(\vec{x}_1 \cdot \vec{x}_2 + R)^d$、$e^{-\frac{||\vec{x}_1 - \vec{x}_2||^2}{2\sigma^2}}$

1.2.0.2 松弛变量

$$\begin{aligned} & \max_{\vec{w},b,\vec{\xi}} \tfrac{1}{||\vec{w}||} - \lambda||\vec{\xi}|| \\ \text{s.t.} \quad & (\vec{w}\cdot\check{x}_\varsigma + b)y_\varsigma \geq 1 - \xi_\varsigma \\ & \xi_\varsigma \geq 0 \end{aligned}$$

等价于

$$\begin{aligned} & \min_{\vec{w}} \tfrac{1}{2}||\vec{w}||^2 + \lambda||\vec{\xi}|| \\ \text{s.t.} \quad & (\vec{w}\cdot\check{x}_\varsigma + b)y_\varsigma \geq 1 - \xi_\varsigma \\ & \xi_\varsigma \geq 0 \end{aligned}$$

等价于

$$\begin{aligned} & L(\vec{w},\vec{c}) = \tfrac{1}{2}||\vec{w}||^2 - \sum_\varsigma c_\varsigma((\vec{w}\cdot\check{x}_\varsigma + b)y_\varsigma - (1-\xi_\varsigma)) + \lambda||\vec{\xi}|| - \sum_\varsigma d_\varsigma\xi_\varsigma \\ & \min_{\vec{w},b,\vec{\xi}} \max_{\vec{c},\vec{d}} L(\vec{w},\vec{c}) \\ \text{s.t.} \quad & c_\varsigma \geq 0 \\ & d_\varsigma \geq 0 \end{aligned}$$

等价于

$$\begin{aligned} & L(\vec{w},\vec{c}) = \tfrac{1}{2}||\vec{w}||^2 - \sum_\varsigma c_\varsigma((\vec{w}\cdot\check{x}_\varsigma + b)y_\varsigma - (1-\xi_\varsigma)) + \lambda||\vec{\xi}|| - \sum_\varsigma d_\varsigma\xi_\varsigma \\ & \max_{\vec{c},\vec{d}} \min_{\vec{w},b,\vec{\xi}} L(\vec{w},\vec{c}) \\ \text{s.t.} \quad & c_\varsigma \geq 0 \\ & d_\varsigma \geq 0 \end{aligned}$$

等价于

(由 $\frac{\partial}{\partial w_i}L(\vec{w},\vec{c}) = 0$ 、 $\frac{\partial}{\partial b}L(\vec{w},\vec{c}) = 0$ 、 $\frac{\partial}{\partial \xi_\varsigma}L(\vec{w},\vec{c}) = 0$ 得：$\vec{w} = \sum_\varsigma c_\varsigma y_\varsigma \check{x}_\varsigma$ 、 $0 = \sum_\varsigma c_\varsigma y_\varsigma$ 、 $\lambda - c_\varsigma = d_\varsigma$)

$$\begin{aligned} & L(\vec{c}) = \sum_\varsigma c_\varsigma - \tfrac{1}{2}\sum_{\varsigma_1}\sum_{\varsigma_2} c_{\varsigma_1}c_{\varsigma_2}y_{\varsigma_1}y_{\varsigma_2}\check{x}_{\varsigma_1}\cdot\check{x}_{\varsigma_2} \\ & \max_{\vec{c}} L(\vec{c}) \\ \text{s.t.} \quad & 0 \leq c_\varsigma \leq \lambda \\ & \sum_\varsigma c_\varsigma y_\varsigma = 0 \end{aligned}$$

（为取$\max\limits_{\vec{c}}$： $\begin{array}{l} \text{当}(\vec{w}\cdot\check{x}_\varsigma + b)y_\varsigma > 1\text{时，} c_\varsigma = 0 \\ \text{当}(\vec{w}\cdot\check{x}_\varsigma + b)y_\varsigma < 1\text{时，} c_\varsigma = C \end{array}$ 。即仅有支持向量和离群向量起作用）

（$f(\vec{x}) = \sum_\varsigma c_\varsigma y_\varsigma \check{x}_\varsigma \cdot \check{x} + b$）

Chapter 2

EM

观测变量x_ς，隐变量u_ς，参数θ

$$P(x_\varsigma|\theta)=\int du_\varsigma P(x_\varsigma,u_\varsigma|\theta)=\int du_\varsigma P(x_\varsigma|u_\varsigma,\theta)P(u_\varsigma|\theta)$$

$$P(\vec{x}|\theta)=\prod_\varsigma P(x_\varsigma|\theta)$$

为求$\theta^*=\mathrm{argmax}_\theta P(\vec{x}|\theta)$：

E: $Q(\theta,\theta^{(t)})\overset{def}{=}\sum\limits_\varsigma E_{P(u_\varsigma|x_\varsigma,\theta^{(t)})}[\log P(x_\varsigma,u_\varsigma|\theta)]=\int du_\varsigma P(u_\varsigma|x_\varsigma,\theta^{(t)})\log P(x_\varsigma,u_\varsigma|\theta)$

M: $\theta^{(t+1)}=\mathrm{argmax}_\theta Q(\theta,\theta^{(t)})$（可放宽为满足$Q(\theta^{(t+1)},\theta^{(t)})>Q(\theta^{(t)},\theta^{(t)})$即可）

反复迭代至收敛（局域最优值）

$$\left(\begin{array}{cl}
 & \log P(\vec{x}|\theta)\\
= & \sum\limits_\varsigma \log P(x_\varsigma|\theta)\\
= & \sum\limits_\varsigma \log\int du_\varsigma P(x_\varsigma,u_\varsigma|\theta)\\
= & \sum\limits_\varsigma \log\int du_\varsigma P(u_\varsigma|x_\varsigma,\theta^{(t)})\frac{P(x_\varsigma,u_\varsigma|\theta)}{P(u_\varsigma|x_\varsigma,\theta^{(t)})}\\
= & \sum\limits_\varsigma \log E_{P(u_\varsigma|x_\varsigma,\theta^{(t)})}[\frac{P(x_\varsigma,u_\varsigma|\theta)}{P(u_\varsigma|x_\varsigma,\theta^{(t)})}]\\
\geq & \sum\limits_\varsigma E_{P(u_\varsigma|x_\varsigma,\theta^{(t)})}[\log\frac{P(x_\varsigma,u_\varsigma|\theta)}{P(u_\varsigma|x_\varsigma,\theta^{(t)})}]\text{（}\theta=\theta^{(t)}\text{时取等号）}\\
 & \theta^{(t+1)}\\
= & \mathrm{argmax}_\theta \log P(\vec{x}|\theta)\\
\simeq & \mathrm{argmax}_\theta \sum\limits_\varsigma E_{P(u_\varsigma|x_\varsigma,\theta^{(t)})}[\log\frac{P(x_\varsigma,u_\varsigma|\theta)}{P(u_\varsigma|x_\varsigma,\theta^{(t)})}]\\
= & \mathrm{argmax}_\theta \sum\limits_\varsigma E_{P(u_\varsigma|x_\varsigma,\theta^{(t)})}[\log P(x_\varsigma,u_\varsigma|\theta)]
\end{array}\right)$$

2.1 GMM(高斯混合模型)

从K个正态分布中挑出一个，挑到第k个概率为α_k，记$u_{\varsigma k}=I$(第ς次挑中第k个正态分布)。第k个正态分布的概率$\phi(x_\varsigma|\mu_k\sigma_k)=\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{(x_\varsigma-\mu_k)^2}{2\sigma_k^2}}$

$P(x_\varsigma|\vec{\alpha}\vec{\mu}\vec{\sigma}) = \prod_k (\alpha_k \phi(x_\varsigma|\mu_k\sigma_k))^{u_{\varsigma k}}$

$P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)}) = \frac{P(x_\varsigma|u_{\varsigma k}, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})P(u_{\varsigma k}|\alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}{\sum\limits_{k'} P(x_\varsigma|u_{\varsigma k'}, \alpha_{k'}^{(t)}\mu_{k'}^{(t)}\sigma_{k'}^{(t)})P(u_{\varsigma k'}|\alpha_{k'}^{(t)}\mu_{k'}^{(t)}\sigma_{k'}^{(t)})} = \frac{\alpha_k^{(t)}\phi(x_\varsigma|u_{\varsigma k}, \mu_k^{(t)}\sigma_k^{(t)})}{\sum\limits_{k'} \alpha_{k'}^{(t)}\phi(x_\varsigma|u_{\varsigma k'}, \mu_{k'}^{(t)}\sigma_{k'}^{(t)})}$

$Q(\theta|\theta^{(t)}) = \sum\limits_\varsigma \sum\limits_k P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})u_{\varsigma k}(\ln \frac{\alpha_k}{\sqrt{2\pi}\sigma_k} - \frac{(x_\varsigma-\mu_k)^2}{2\sigma_k^2})$

反复迭代

- $\mu_k^{(t+1)} = \frac{\sum\limits_\varsigma x_\varsigma u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}{\sum\limits_\varsigma u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}$
- $\sigma_k^{(t+1)} = \sqrt{\frac{\sum\limits_\varsigma (x_\varsigma-\mu_\varsigma)^2 u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}{\sum\limits_\varsigma u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}}$
- $\alpha_k^{(t+1)} = \frac{\sum\limits_\varsigma u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}{\sum\limits_{k'}\sum\limits_\varsigma u_{\varsigma k'} P(u_{\varsigma k'}|x_\varsigma, \alpha_{k'}^{(t)}\mu_{k'}^{(t)}\sigma_{k'}^{(t)})}$

2.2 K-means

将训练数据集聚类为K类

$\alpha_k = \frac{1}{K}$，$\sigma_k = 1$

$P(x_\varsigma|\vec{\mu}) = \frac{1}{K}\prod_k (\phi(x_\varsigma|\mu_k))^{u_{\varsigma k}}$

$P(u_{\varsigma k}|x_\varsigma, \mu_k^{(t)}) = \frac{P(x_\varsigma|u_{\varsigma k}, \mu_k^{(t)})P(u_{\varsigma k}|\mu_k^{(t)})}{\sum\limits_{k'} P(x_\varsigma|u_{\varsigma k'}, \mu_{k'}^{(t)})P(u_{\varsigma k'}|\mu_{k'}^{(t)})} = \frac{\phi(x_\varsigma|u_{\varsigma k}, \mu_k^{(t)})}{\sum\limits_{k'} \phi(x_\varsigma|u_{\varsigma k'}, \mu_{k'}^{(t)})}$

$Q(\theta|\theta^{(t)}) = \sum\limits_\varsigma \sum\limits_k P(u_{\varsigma k}|x_\varsigma, \mu_k^{(t)})u_{\varsigma k}(\ln \frac{1}{\sqrt{2\pi}\sigma_k K} - \frac{1}{2}(x_\varsigma - \mu_k)^2)$

反复迭代

- $\mu_k^{(t+1)} = \frac{\sum\limits_\varsigma x_\varsigma u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}{\sum\limits_\varsigma u_{\varsigma k} P(u_{\varsigma k}|x_\varsigma, \alpha_k^{(t)}\mu_k^{(t)}\sigma_k^{(t)})}$

为简化计算，取$P(u_{\varsigma k}|x_\varsigma, \mu_k^{(t)}) \simeq I($ x_ς 离$\mu_k^{(t)}$ 最近)，即将每个x_ς归到离它最近的$\mu_k^{(t)}$类上

则$\mu_k^{(t+1)} =$ 所有第k类的x_ς的平均值

Chapter 3

流型学习

分布在低维流型上的样本$\{\vec{y}_\varsigma\}$ 被光滑嵌入f嵌入到高维空间中，在观察到高维空间中样本$\{\vec{x}_\varsigma\}$的条件下重构f与$\{\vec{y}_\varsigma\}$

3.1 PCA（主成分分析）

无监督学习。为找出最能代表训练样本$\{\vec{x}_\varsigma\}$的方向，即求

$$\forall ij \quad \begin{array}{rl} & \max\limits_{||w_i||=1} Var[\vec{w}_i X] \\ s.t. & Cov[\vec{w}_i X, \vec{w}_j X] = 0 \end{array}$$

令$\check{x}'_\varsigma \stackrel{def}{=} \check{x}_\varsigma - E[\check{x}]$

假设SVD分解得：

$$\begin{bmatrix} \cdots & \check{x}'_\varsigma & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix} \begin{bmatrix} \ddots & & \\ & \lambda_k & \\ & & \ddots \end{bmatrix} V^T$$

则将$\check{x}'$投影到各$\vec{w}_k$方向上：

$$\vec{z}'^T = \check{x}'^T \begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix}$$

只取特征值λ_k最大的若干个$\vec{w}_k$ （即只取$\begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix}$的左半段矩阵）作为特征方向进行投影，能最大限度分离各$\check{x}$

3.1.1 法一

矩阵

$$\begin{bmatrix} \cdots & \check{x}'_\varsigma & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \check{x}'^T_\varsigma \\ \vdots \end{bmatrix} = \begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix} \begin{bmatrix} \ddots & & \\ & \lambda_k^2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ \vec{w}_k^T \\ \vdots \end{bmatrix}$$

特征值与特征向量为$\{(\vec{w}, \lambda_k^2)\}$

3.1.2 法二

将$\check{x}'_\varsigma$投影到各$\vec{w}_k$方向上：

$$\begin{bmatrix} \vdots \\ \check{z}'^T_\varsigma \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \check{x}'^T_\varsigma \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix} = V \begin{bmatrix} \ddots & & \\ & \lambda_k & \\ & & \ddots \end{bmatrix}$$

则

$$\begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \check{x}_\varsigma & \cdots \end{bmatrix} V \begin{bmatrix} \ddots & & \\ & \frac{1}{\lambda_k} & \\ & & \ddots \end{bmatrix} = \begin{bmatrix} \cdots & \check{x}'_\varsigma & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \check{z}'^T_\varsigma \\ \vdots \end{bmatrix} \begin{bmatrix} \ddots & & \\ & \frac{1}{\lambda_k^2} & \\ & & \ddots \end{bmatrix}$$

矩阵

$$\begin{bmatrix} \vdots \\ \check{x}'^T_\varsigma \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \check{x}'_\eta & \cdots \end{bmatrix} = V \begin{bmatrix} \ddots & & \\ & \lambda_k^2 & \\ & & \ddots \end{bmatrix} V^T = \begin{bmatrix} \vdots \\ \check{z}'^T_\varsigma \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \check{z}'_\eta & \cdots \end{bmatrix}$$

的特征向量与特征值为$\{(\frac{1}{\lambda_k} \begin{bmatrix} \vdots \\ z'_{\varsigma k} \\ \vdots \end{bmatrix}, \lambda_k^2)\}$

则

$$\check{z}'^T = \check{x}'^T \begin{bmatrix} \cdots & \vec{w}_k & \cdots \end{bmatrix} = \check{x}'^T \begin{bmatrix} \cdots & \check{x}'_\varsigma & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \check{z}'^T_\varsigma \\ \vdots \end{bmatrix} \begin{bmatrix} \ddots & & \\ & \frac{1}{\lambda_k^2} & \\ & & \ddots \end{bmatrix}$$

3.1.3 Kernel PCA

当$\check{x}$太高维甚至无穷维时，无法显式求出$\vec{w}_k$
则可在“法二”中，替换

$$\begin{bmatrix} \vdots \\ \check{x}'^T_\varsigma \\ \vdots \end{bmatrix} \begin{bmatrix} \cdots & \check{x}'_\eta & \cdots \end{bmatrix} = \begin{bmatrix} & \vdots & \\ \cdots & K(\vec{x}'_\varsigma, \vec{x}'_\eta) & \cdots \\ & \vdots & \end{bmatrix}$$

$$\check{x}'^T \begin{bmatrix} \cdots & \check{x}'_\varsigma & \cdots \end{bmatrix} = \begin{bmatrix} & \vdots & \\ \cdots & K(\vec{x}', \vec{x}'_\varsigma) & \cdots \\ & \vdots & \end{bmatrix}$$

3.1.4 白化

PCA后可进行白化，对$\check{z}_k$ 均值归零，方差归一

3.2 MDS（多维度尺度变换）

无监督学习。已知任意两点间距$\delta_{\varsigma\eta}$。重构向量$\vec{x}_\varsigma$，使得各点间距为$\delta_{\varsigma\eta}$
即求

$$\min_{\vec{x}_1,\cdot,\vec{x}_{\mathbb{N}}}\sum_{\varsigma\eta}(||\vec{x}_\varsigma-\vec{x}_\eta||-\delta\varsigma\eta)^2$$

定义内积$t_{\varsigma\eta}\stackrel{def}{=}(\vec{x}_\varsigma-E[\vec{x}])\cdot(\vec{x}_\eta-E[\vec{x}])$，距离$d_{\varsigma\eta}\stackrel{def}{=}||\vec{x}_\varsigma-\vec{x}_\eta||$，则内积

$$t_{\varsigma\eta}=-\frac{1}{2}\left(d_{\varsigma\eta}^2-\frac{1}{\mathbb{N}}\sum_\mu d_{\varsigma\mu}^2-\frac{1}{\mathbb{N}}\sum_\nu d_{\nu\eta}^2+\frac{1}{\mathbb{N}^2}\sum_{\mu\nu}d_{\mu\nu}^2\right)$$

可完全用距离$d_{\varsigma\eta}$表示
分解

$$\begin{bmatrix}\vdots\\ \vec{x}_\varsigma'^T\\ \vdots\end{bmatrix}\begin{bmatrix}\cdots & \vec{x}_\eta' & \cdots\end{bmatrix}=\begin{bmatrix} & \vdots & \\ \cdots & t_{\varsigma\eta} & \cdots\\ & \vdots & \end{bmatrix}=U^T\Lambda U=(\Lambda^{\frac{1}{2}}U)^T(\Lambda^{\frac{1}{2}}U)$$

取特征值最大的k'个$\vec{u}_i$分量作为$\vec{x}$

3.3 isomap

无监督学习。映射过程中尽可能保持全局流形上测地线的距离
未知流型结构，用有限数据采样估计测地线：
构造邻接图$W_{ij}=\begin{cases}1 & , \quad ||\vec{x}_i-\vec{x}_j||<\varepsilon\\ 0 & , \quad ||\vec{x}_i-\vec{x}_j||\geq\varepsilon\end{cases}$，则任意两点间测地线$d_M$用邻接图上的最短路径长度近似
用MDS 计算映射后的坐标$\vec{y}$，使得映射坐标下的欧氏距离与原来的测地线距离尽量相等：

$$\min_{\vec{y}}\sum_{i,j}(||\vec{y}_i-\vec{y}_j||-d_M(\vec{x}_i,\vec{x}_j))^2$$

3.4 LLE

无监督学习。由流型在局部等价于欧几里得空间，尽可能保持流型局部线性关系
对任意点$\vec{x}_\varsigma$，只考虑其周围的点$\vec{x}_\eta$（记为$\eta\sim\varsigma$）：

1. 将高维坐标间关系反映到权重w中：$\mathrm{argmin}_w\sum_\varsigma||\vec{x}_\varsigma-\sum_{\eta\sim\varsigma}w_{\varsigma\eta}\vec{x}_\eta||^2$
2. 将权重w反映到低维坐标$\vec{y}$中：$\mathrm{argmin}_{\vec{y}}\sum_\varsigma||\vec{y}_\varsigma-\sum_{\eta\sim\varsigma}w_{\varsigma\eta}\vec{y}_\eta||^2$

3.5 LDA（线性判别分析）

监督学习，分类。使得投影后类内方差最小，类间方差最大
训练样本集$\{(\check{x}_\varsigma,y_\varsigma)\}$，其中$y_\varsigma$属于有限的离散值（分类问题）

- 整体散度$S_T \overset{def}{=} \sum_{\check{x}_\varsigma \in D} (\check{x}_\varsigma - \bar{\check{x}})(\check{x}_\varsigma - \bar{\check{x}})^T$
- 类内散度$S_W \overset{def}{=} \sum_i \sum_{\check{x}_\varsigma \in D_i} (\check{x}_\varsigma - \bar{\check{x}}_i)(\check{x}_\varsigma - \bar{\check{x}}_i)^T$
- 类间散度$S_B \overset{def}{=} S_T - S_W = \sum_i \mathbb{N}_i (\bar{\check{x}}_i - \bar{\check{x}})(\bar{\check{x}}_i - \bar{\check{x}})^T$

投影到一维$\vec{w}$上，$z_\varsigma = \vec{w}^T \check{x}_\varsigma$
则目标函数为投影后的

$$\max_{\vec{w}} \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}$$

等价于（因$\vec{w}$的模长不重要）

$$\begin{aligned} &\max_{\vec{w}} \vec{w}^T S_B \vec{w} \\ s.t. \quad &\vec{w}^T S_W \vec{w} = 1 \end{aligned}$$

Chapter 4

ANN

4.1 CNN（卷积神经网络）

4.1.1 卷积层

提取特征。

第l层、第k个卷积核：$N^{l,k}*N^{l,k}$的卷积核与输入图层每$N^{l,k}*N^{l,k}$的框点乘。框之间有重叠。

$$z^{l,k}_{m,n} = f(w^{l,k}x_{i,j} + b^{l,k}) \quad (i,j \in m,n \pm \frac{N^{l,k}-1}{2})$$

4.1.2 池化层

平移对称性。

第l层：输入图层每$N^{l,k}*N^{l,k}$的框选出代表值。框之间不重叠

$$z^{l}_{m,n} = \mathrm{pool}(x_{i,j}) \quad (i,j \in m,n \pm \frac{N^{l,k}-1}{2})$$

4.1.3 全连接层

4.2 RNN

4.2.1 RNN

输入单元$\{\cdots, x_{t-1}, x_t, x_{t+1}, \cdots\}$，隐藏单元$\{\cdots, s_{t-1}, s_t, s_{t+1}, \cdots\}$，输出单元$\{\cdots, y_{t-1}, y_t, y_{t+1}, \cdots\}$。

$$s_t = f(Ux_t + Ws_{t-1})$$

$$o_t = softmax(Vs_t)$$

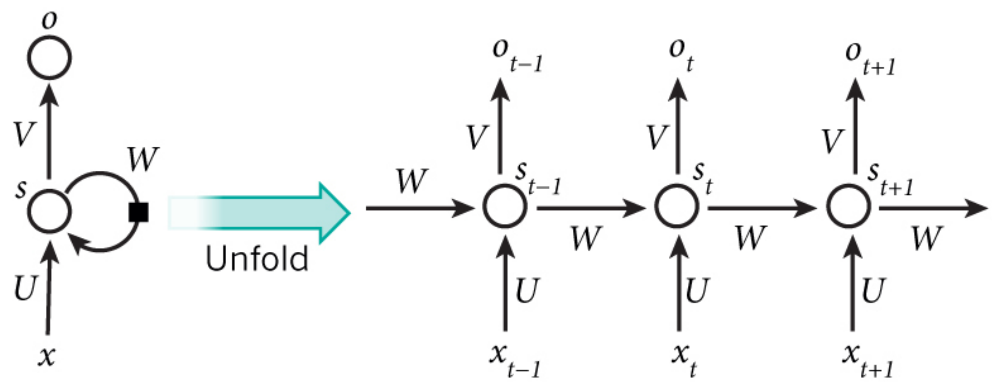

Figure 4.1: RNN

4.2.2 LSTM

4.2.2.1 LSTM

$$\begin{aligned}
\text{记忆门}_t &= \sigma(W_f * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{输入门}_t &= \sigma(W_i * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{新值}_t &= tanh(W_c * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{状态}_t &= \text{记忆门}_t * \text{状态}_{t-1} + \text{输入门}_t * \text{新值}_t \\
\text{输出门}_t &= \sigma(W_o * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{输出}_t &= \text{输出门} * tanh(\text{状态}_t)
\end{aligned}$$

4.2.2.2 peephole connection

$$\begin{aligned}
\text{记忆门}_t &= \sigma(W_f * [\text{状态}_{t-1}, \text{输出}_{t-1}, \text{输入}_t]) \\
\text{输入门}_t &= \sigma(W_i * [\text{状态}_{t-1}, \text{输出}_{t-1}, \text{输入}_t]) \\
\text{新值}_t &= tanh(W_c * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{状态}_t &= \text{记忆门}_t * \text{状态}_{t-1} + \text{输入门}_t * \text{新值}_t \\
\text{输出门}_t &= \sigma(W_o * [\text{状态}_t, \text{输出}_{t-1}, \text{输入}_t]) \\
\text{输出}_t &= \text{输出门} * tanh(\text{状态}_t)
\end{aligned}$$

4.2.2.3 coupled记忆门与输入门

$$\begin{aligned}
\text{记忆门}_t &= \sigma(W_f * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{输入门}_t &= 1 - \text{记忆门}_t \\
\text{新值}_t &= tanh(W_c * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{状态}_t &= \text{记忆门}_t * \text{状态}_{t-1} + \text{输入门}_t * \text{新值}_t \\
\text{输出门}_t &= \sigma(W_o * [\text{输出}_{t-1}, \text{输入}_t]) \\
\text{输出}_t &= \text{输出门} * tanh(\text{状态}_t)
\end{aligned}$$

4.2.2.4 GRU(Gated Recurrent Unit)

$$
\begin{aligned}
\text{记忆门}_t &= 1-\text{输入门}_t \\
\text{输入门}_t &= \sigma(W_i * [\text{状态}_{t-1}, \text{输出}_{t-1}, \text{输入}_t]) \\
\text{R门}_t &= \sigma(W_r * [\text{状态}_{t-1}, \text{输出}_{t-1}, \text{输入}_t]) \\
\text{新值}_t &= tanh(W_c * [\text{R门}_t * \text{输出}_{t-1}, \text{输入}_t]) \\
\text{状态}_t &= \text{记忆门}_t * \text{状态}_{t-1} + \text{输入门}_t * \text{新值}_t \\
\text{输出门}_t &= \text{无} \\
\text{输出}_t &= \text{状态}_t
\end{aligned}
$$

4.3 AE（自编码）

4.3.1 AE

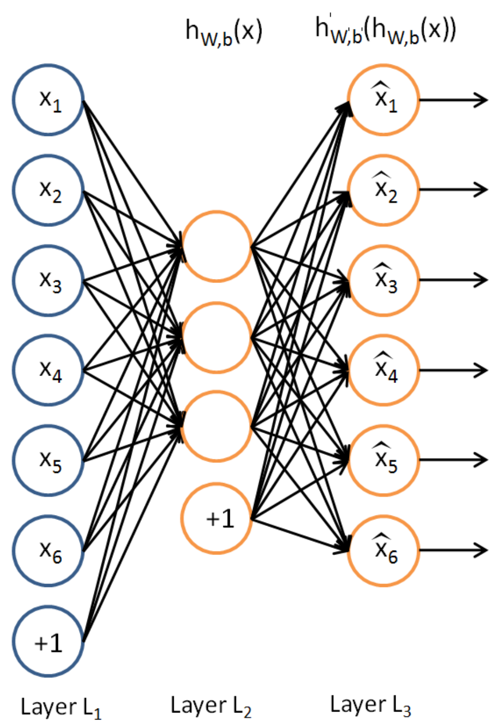

Figure 4.2: AE

一层隐藏层+一层输出层

无监督学习。输出层尽力还原输入层，则中间隐藏层为提取的特征。

- 隐藏层可取sigmoid函数
- 输出层取线性函数时，可取$L(x, \hat{x}) = \frac{1}{2}||x-\hat{x}||$

  输出层取sigmoid函数时，可取$L(x, \hat{x}) = -\sum_i (x_i \log \hat{x}_i + (1-x_i)\log(1-\hat{x}_i))$

4.3.2 稀疏性限制

限制神经元大部分时间$\vec{w} \cdot \check{x} < 0$

神经元j的激活度$\rho_j = \frac{1}{\mathbb{N}} \sum_{\varsigma=1}^{\mathbb{N}} [f(\vec{w}_j \cdot \check{x_\varsigma})]$，期望接近于一个特定值$\rho$（譬如f为sigmoid函数时，可取$\rho = 0.05$）

在优化目标函数中加入惩罚因子$\sum_{j \in 隐藏层} KL(\rho || \rho_j)$。其中相对熵$KL(\rho||\rho_j) = \rho \ln \frac{\rho}{\rho_j} + (1-\rho) \ln \frac{1-\rho}{1-\rho_j}$

4.3.3 Denoising Auto-Encoder

为提高鲁棒性，在自编码模型中，将输入x添加破坏变为y，经过自编码器得到$\hat{y}$，目标优化函数为$L(x, \hat{y})$。

可取$y = x +$高斯模型，或直接将x的某些分量随机为0得到y。

4.3.4 多层AE

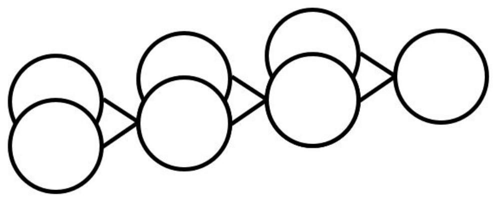

Figure 4.3: MulAE

逐层训练，每一层提取的特征作为下一层输入。（训练完毕后再进行有监督训练为早期深度学习做法）

4.4 ELM（超限学习机）

一层隐藏层+一个输出结点

训练样本集$\{(\vec{x}_\varsigma, \vec{y}_\varsigma)\}$。

隐藏层L个结点输出

$$\begin{bmatrix} & \vdots & \\ \cdots & f(\vec{W}_l \cdot \check{x}_\varsigma) & \cdots \\ & \vdots & \end{bmatrix}$$

输出层输出

$$\begin{bmatrix} \cdots & \vec{\beta}_l & \cdots \end{bmatrix} \begin{bmatrix} & \vdots & \\ \cdots & f(\vec{W}_l \cdot \check{x}_\varsigma) & \cdots \\ & \vdots & \end{bmatrix} = \begin{bmatrix} \cdots & \vec{y}_\varsigma & \cdots \end{bmatrix}$$

随机取$\vec{W}_l$，固定不变。则

$$\begin{bmatrix} \cdots & \vec{\beta}_l & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \vec{y}_\varsigma & \cdots \end{bmatrix} \begin{bmatrix} & \vdots & \\ \cdots & f(\vec{W}_l \cdot \check{x}_\varsigma) & \cdots \\ & \vdots & \end{bmatrix}^+$$

其中 $^+$为Moore-Penrose广义逆

4.5 Hopfield

4.5.1 Hopfield

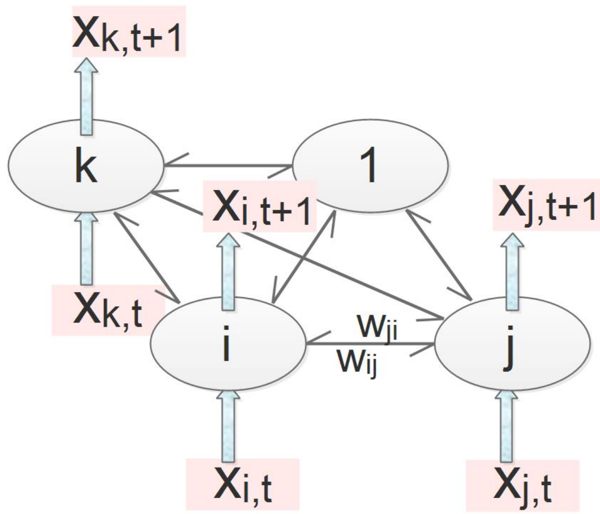

Figure 4.4: Hopfield

训练时：

$$E(\check{x}) \overset{def}{=} -\frac{1}{2}\begin{bmatrix} \cdots & x_i & \cdots \end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ij} & \cdots \\ & \vdots & \end{bmatrix}\begin{bmatrix} \vdots \\ x_j \\ \vdots \end{bmatrix}$$

可取$\begin{cases} w_{ij} = w_{ji} \\ w_{ii} = 0 \end{cases}$，$x_i$只取双值。保证能量有最小值

优化方法：

1. 保持$\vec{w}$不变，改变$\check{x}$，至能量最小
2. 保持$\check{x}$不变，x_i、x_j值相同则增大w_{ij}，x_i、x_j值不同则减小w_{ij}
（譬如当x_i双值为1,-1时，$w_{ij} = \sum_{\varsigma} x_{\varsigma i}x_{\varsigma j}$）

预测时：

4.5.1.1 DHNN（离散时间）

$$\begin{bmatrix} \vdots \\ x_{i,t+1} \\ \vdots \end{bmatrix} = f(\begin{bmatrix} & \vdots & \\ \cdots & w_{ij} & \cdots \\ & \vdots & \end{bmatrix}\begin{bmatrix} \vdots \\ x_{j,t} \\ \vdots \end{bmatrix})$$

初始输入$\check{x}_0$进行迭代收敛至稳定点$\check{x}_T$，用$\check{x}_T$进行判别

可取

$$f(z)=\begin{cases}-1 & , \quad z<0\\ 1 & , \quad z\geq 0\end{cases}$$

或

$$f(z)=\begin{cases}-1 & , \quad z<-1\\ z & , \quad -1\leq z\leq 1\\ 1 & , \quad z>1\end{cases}$$

4.5.1.2 CHNN（连续时间）

$$E(\check{x})\overset{def}{=}-\frac{1}{2}\begin{bmatrix}\cdots & x_i & \cdots\end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ij} & \cdots\\ & \vdots & \end{bmatrix}\begin{bmatrix}\vdots\\ x_j\\ \vdots\end{bmatrix}+\begin{bmatrix}\cdots & \frac{1}{R_i} & \cdots\end{bmatrix}\begin{bmatrix}\vdots\\ \int\limits_0^{x_i} f^{-1}(x')dx'\\ \vdots\end{bmatrix}$$

$$E(\check{x}(t))\overset{def}{=}-\frac{1}{2}\begin{bmatrix}\cdots & x_i(t) & \cdots\end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ij} & \cdots\\ & \vdots & \end{bmatrix}\begin{bmatrix}\vdots\\ x_j(t)\\ \vdots\end{bmatrix}+\begin{bmatrix}\cdots & \frac{1}{R_i} & \cdots\end{bmatrix}\begin{bmatrix}\vdots\\ \int\limits_0^{x_i(t)} f^{-1}(x')dx'\\ \vdots\end{bmatrix}$$

可取

$$f(z)=\frac{1}{1-e^{-z}}$$

4.5.2 Ising模型

$$w_{ij}=\begin{cases}J & \text{ij近邻}\\ H & \text{ij其中一个x为1}\\ 0 & \text{ij非近邻}\end{cases}$$

微观构型$\check{x}$的概率$p(\check{x})=\frac{1}{Z}e^{-\frac{E(\check{x})}{kT}}$ （配分函数$Z=\sum\limits_{\check{x}}e^{-\frac{E(\check{x})}{kT}}$）

4.5.3 RBM（受限波尔兹曼机）

一层显层+一层隐层

能量

$$E(\check{v},\check{h})\overset{def}{=}-\begin{bmatrix}\cdots & v_i & \cdots\end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ij} & \cdots\\ & \vdots & \end{bmatrix}\begin{bmatrix}\vdots\\ h_j\\ \vdots\end{bmatrix}$$

概率

$$P(\check{v},\check{h})\overset{def}{=}\frac{1}{Z}e^{-E(\check{v},\check{h})}$$

自由能

$$F(\check{v})\overset{def}{=}-ln\sum_h e^{-E(\check{v},\check{h})}$$

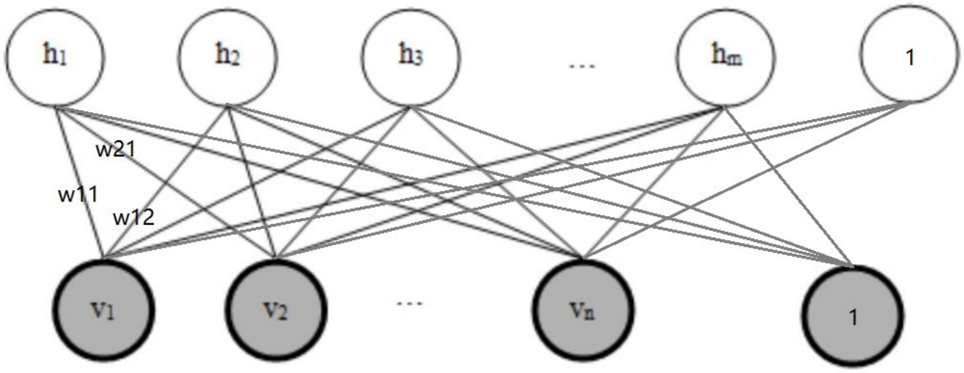

Figure 4.5: RBM

$$P(\check{v}) = \sum_{h} P(\check{v}, \check{h}) = \frac{1}{Z} e^{-F(\check{v})}$$

优化目标

$$\text{argmax}_w \prod_{\check{v}_\varsigma} P(\check{v}) = \text{argmin}_w \sum_{\check{v}_\varsigma} F(\check{v})$$

即提取显层（训练样本）的特征藏于隐层参数中，最大概率还原显层

4.5.3.1 对比散度训练法

由每一个训练样本$\check{v}$ 求得$\check{h}$，再由$\check{h}$ 反求得$\check{v}'$。则$w_{ij}+ = \lambda(v_i - v_i')h_j$ 。循环训练至收敛$(x_i = x_i')$。
改进：

1. 用多往返几次的$\check{v}'''$替代$\check{v}'$。
2. 用$p(v_i)$、$p(h_j)$替代v_i、h_j
3. 加正则项，对较大的权重w_{ij}进行惩罚
4. 用本次Δw_{ij}与多次前次Δw_{ij}线性加权

4.5.3.2 二项分布

假设h_j、v_i都只能取$\{0,1\}$

$$\begin{bmatrix} \vdots \\ p(h_j = 1|\check{v}) \\ \vdots \end{bmatrix} = \sigma\left(\begin{bmatrix} & \vdots & \\ \cdots & w_{ji} & \cdots \\ & \vdots & \end{bmatrix}\begin{bmatrix} \vdots \\ v_i \\ \vdots \end{bmatrix}\right)$$

$$\begin{bmatrix} \cdots & p(v_i = 1|\check{h}) & \cdots \end{bmatrix} = \sigma\left(\begin{bmatrix} \cdots & h_j & \cdots \end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ji} & \cdots \\ & \vdots & \end{bmatrix}\right)$$

4.5.3.3 多项分布

假设h_j、v_i都只能取$(0,\cdots,0,1,0,\cdots,0)$其中之一

$$\begin{bmatrix} \cdots & p(v_i^k=1|\check{h}) & \cdots \end{bmatrix} = \frac{\exp\left(\begin{bmatrix} \cdots & h_j & \cdots \end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ji}^k & \cdots \\ & \vdots & \end{bmatrix}\right)}{\sum_{k=1}^{K}\exp\left(\begin{bmatrix} \cdots & h_j & \cdots \end{bmatrix}\begin{bmatrix} & \vdots & \\ \cdots & w_{ji}^k & \cdots \\ & \vdots & \end{bmatrix}\right)}$$

4.5.4 DBN（深度信念网络）

多层RBM组成，逐层训练，每层提取的特征作为下一层输入（即当前隐层作为下层隐层)。（训练完毕后再进行有监督训练为早期深度学习做法）

4.6 Highway Network

4.6.1 Highway Network[1] [2]

将一层或多层由原本的$\check{y}=F(\check{x})$ 改为$\check{y}=F(\check{x})*T(W_T\check{x})+W_s\check{x}*C(W_C\check{x})$
（$*$表示按元素乘）
（W_s用于将$\check{x}$的维度转为与$F(W_F\check{x})$一致）

4.6.2 Deep Residual Network[3]

若某一多层网络可渐进估计某函数$H(\check{x})$，则等同可渐进估计$H(\check{x})-\check{x}$
$W_T=T=W_C=C=1$，即$\check{y}=F(\check{x})+W_s\check{x}$

4.7 GAN（生成对抗网络）

非监督学习
生成器网络G：由真训练样本集生成假样本。判别器网络D：辨别样本真假（二分类器）

$$\min_G\max_D\left(E_{x\sim P_{\text{data}}}[\log D(x)]+E_{z\sim P_{\text{noise}}}[\log(1-D(G(z)))]\right)$$

最终生成器网络与判别器网络达到纳什均衡。生成器完美复原训练数据分布，判别器准确率为50
训练过程中固定一方，更新另一方参数，交替迭代，使对方错误最大化

4.7.1 CGAN（条件生成式对抗网络）

加入监督信息y作为条件，进行约束

$$\min_G\max_D\left(E_{x\sim P_{\text{data}}}[\log D(x|y)]+E_{z\sim P_{\text{noise}}}[\log(1-D(G(z|y)))]\right)$$

4.8 其他

4.8.1 Dropout

训练过程中，每个神经元以一定概率p失活为0，若非0则其输出结果再除以p以恢复原大小
预测时不失活

4.8.2 Batch Normalization

在每一层网络前都加入一层数据处理。
为节省时间，不投影到特征上。
将x_j 化为均值0方差1：$x_j'' = \frac{x_j - \mu_j}{\sqrt{\sigma_j^2 + \varepsilon}}$
将x_j'' 进行变换：$x_j''' = \gamma x_j'' + \beta$。$\gamma$、$\beta$作为参数在网络中迭代训练

Chapter 5

强化学习

5.1 强化学习

5.1.1 基本概念

状态s，动作a

学习策略：当前s下采取a的概率$\pi(s,a)$

系统反馈：当前s_1下采取a后变为s_2的概率$P(s_1 \xrightarrow{a} s_2)$

奖励$R_{s,a}$，衰减因子γ

状态-动作价值

$$\begin{aligned} q_\pi(s,a) &\stackrel{def}{=} E_\pi[\textstyle\sum_{k=0}^{\infty}\gamma^k R_{k+t}|s_t=s,a_t=a] \\ &= E_\pi[R_t+\gamma q_\pi(s_{t+1},a_{t+1})|s_t=s,a_t=a] \\ &= R_{s,a}+\gamma\textstyle\sum_{s'\in S}P(s\xrightarrow{a}s')\sum_{a'\in A}\pi(s',a')q(s',a') \\ &= R_{s,a}+\gamma\textstyle\sum_{s'\in S}P(s\xrightarrow{a}s')v(s') \end{aligned}$$

状态价值

$$\begin{aligned} v_\pi(s) &\stackrel{def}{=} E_\pi[\textstyle\sum_{k=0}^{\infty}\gamma^k R_{k+t}|s_t=s] \\ &= E_\pi[R_t+\gamma v_\pi(s_{t+1})|s_t=s] \\ &= \textstyle\sum_{a\in A}\pi(s,a)\left(R_{s,a}+\gamma\sum_{s'\in S}P(s\xrightarrow{a}s')v(s')\right) \\ &= \textstyle\sum_{a\in A}\pi(s,a)q(s,a) \end{aligned}$$

5.1.2 已知模型

已知$R_{s,a}$，$P(s_1 \xrightarrow{a} s_2)$ 下的学习

状态价值：

$$v(s)=\sum_{a\in A}\pi(s,a)[R_{s,a}+\gamma\sum_{s'\in S}P(s\xrightarrow{a}s')v(s')]$$

或

$$v(s)=\max_a[R_{s,a}+\gamma\sum_{s'\in S}P(s\xrightarrow{a}s')v(s')]$$

迭代$v(s)$至收敛

更新策略：

$$\pi(s,a)=\begin{cases}1-\varepsilon & ,\quad (a=\text{argmax}_a q(s,a))\\ \frac{\varepsilon}{|A-1|} & ,\quad (a\neq \text{argmax}_a q(s,a))\end{cases}$$

循环以上两步至收敛

5.1.3 未知模型

$R_{s,a}$，$P(s_1 \xrightarrow{a} s_2)$ 未知，仅由环境反馈s、R下的学习：

5.1.3.1 蒙特卡洛法

由当前π生成序列：$s_1,a_1,R_1,\cdots,s_k,a_k,R_k$

每个时刻t，

$$\begin{array}{lll}\text{每次抽样} & \begin{cases}V(s_t)+=\sum_{i=0}^{k-t}\gamma^i R_{t+i}\\ ++N(s_t)\end{cases} & \begin{cases}Q(s_t,a)+=\sum_{i=0}^{k-t}\gamma^i R_{t+i}\\ ++N(s_t,a_t)\end{cases}\\ \text{最终} & v(s_t)=\frac{V(s_t)}{N(s_t)} & q(s_t,a_t)=\frac{Q(s_t,a_t)}{N(s_t,a_t)}\end{array}$$

5.1.3.2 时差学习

TD(0)

$$\begin{array}{lllll}v(s) & = & \lambda(r+\gamma v(s')) & + & (1-\lambda)v(s)\\ q(s,a) & = & \lambda(r+\gamma q(s',a')) & + & (1-\lambda)q(s,a)\\ q(s,a) & = & \lambda(r+\gamma max_{a'}q(s',a')) & + & (1-\lambda)q(s,a)\end{array}$$

5.2 DQN（深度强化学习）

强化学习中，状态s为天文数字，无法构建完整的表$q(s,a)$。故设法用函数$q(s,a,\theta)$拟合$q(s,a)$，用神经网络表示该函数

用Q-learning，逼近

$$q(s,a,\theta)=r+\gamma\max_{a'}q(s',a',\theta)$$

则损失函数

$$L(\theta)\overset{def}{=}E[(r+\gamma\max_{a'}q(s',a',\theta)-q(s,a,\theta))^2]$$

5.2.0.3 Experience Replay

按策略π 生成序列$s_1,a_1,R_1,\cdots,s_k,a_k,R_k$，从中随机抽取若干个进行训练（避免按连续选取会有相干性）。

重复以上若干遍至训练出正确网络$q(s,a,w)$用以拟合$q(s,a)$。

5.2.0.4 Target Q

新旧两个网络。用旧网络进行计算，参数更新至新网络上，延迟一段时间后再将新网络参数复制回旧网络。避免相关性太大。

$$L(\theta) \overset{def}{=} E[(r + \gamma \max_{a'} q(s', a', \theta_{\text{old}}) - q(s, a, \theta_{\text{new}}))^2]$$

5.2.0.5 Double DQN

$$L(\theta) \overset{def}{=} E[(r + \gamma q(s', \text{argmax}_{a'} q(s, a, \theta_{\text{new}}), \theta_{\text{old}}) - q(s, a, \theta_{\text{new}}))^2]$$

5.2.0.6 Prioritised replay

从Experience Replay中抽取(s,a)进行训练时，抽样概率与$|r + \gamma \max_{a'} q(s', a', \theta_{\text{old}}) - q(s, a, \theta_{\text{new}})|$ 成正比。

Chapter 6

决策树

回归树：每个节点都有预测值。最小化均方差，使分到节点中的数据与预测值方差最小
分类树：最大熵

6.1 单决策树

6.1.1 ID3

6.1.1.1 定义

对集合G，属性A将其分为子集G_a（不同G_a有不同A值）
信息熵

$$S(G,A) \stackrel{def}{=} -\sum_a \frac{|G_a|}{|G|} \log \frac{|G_a|}{|G|}$$

信息增益

$$Gain(G,A) \stackrel{def}{=} S(G,\text{正反例}) - \sum_a \frac{|G_a|}{|G|} S(G_a,\text{正反例})$$

6.1.1.2 算法

树各节点为样本集合
对每个节点选取信息增益最大的属性A，该节点中样本对A的不同值生成不同子节点。
持续分类直至每个节点正反取值一致，或用光所有属性。

6.1.2 C4.5

6.1.2.1 定义

信息增益率

$$GainR(G,A) \stackrel{def}{=} \frac{Gain(G,A)}{S(G,A)}$$

6.1.2.2 算法

树各节点为样本集合

对每个节点选取信息增益率最大的属性A，该节点中样本对A的不同值生成不同子节点。

持续分类直至每个节点正反取值一致，或用光所有属性。

6.1.3 最小二乘回归树

空间D划分为多个区域D_s，寻找划分方式S

$$\min_S \left\{ \sum_s \sum_{(x_\varsigma, y_\varsigma) \in D_s} (y_\varsigma - \bar{y}_s)^2 \right\}$$

其中区域D_s的输出值$\bar{y}_s = \frac{1}{|D_s|} \sum_{(x_\varsigma, y_\varsigma) \in D_s} y_\varsigma$

依次递归划分区域

6.1.4 Cart分类树

空间D中，属于第k类的空间$D_k = D \bigcap C_k$，则基尼系数

$$Gini(D) \stackrel{def}{=} \sum_k \frac{|D_k|}{|D|} \left(1 - \frac{|D_k|}{|D|}\right) = 1 - \sum_k \left(\frac{|D_k|}{|D|}\right)^2$$

空间D划分为多个区域D_s，寻找划分方式S

$$\min_S \left\{ \sum_s \frac{|D_s|}{|D|} Gini(D_s) \right\}$$

依次递归划分区域

6.2 Boosting

6.2.1 随机森林

对每棵树，从A个总训练样本中有放回抽取a个作为其训练样本（可取a=A）。

对每个结点，从F个维度属性中不放回抽取f个作为其判断属性，从f个判断属性中找出最佳属性进行划分。

预测时用所有树共同决定分类。

6.2.2 AdaBoost

6.2.2.1 原理

多个弱分类器共同决定分类。

分类错误的训练样本权重加大，分类正确的训练样本权重减小。

训练完毕后，误差率大的弱分类器投票权重较小，误差率小的弱分类器投票权重较大。

6.2.2.2 具体算法

第t轮训练样本$(x_\varsigma, y_\varsigma)$ 的权重为$w_{t,\varsigma}$，构建弱分类器$f_t(x)$ 使分类误差率

$$\varepsilon_t = \sum_\varsigma w_{t,\varsigma} I(f_t(x_\varsigma) \neq y_\varsigma)$$

最小。

分类器f_t的重要程度

$$c_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$$

更新

$$w_{t+1,\varsigma} = \begin{cases} \frac{1}{Z_t} w_{t,\varsigma} e^{-c_t} & , \quad y_\varsigma = f_t(x_\varsigma) \\ \frac{1}{Z_t} w_{t,\varsigma} e^{+c_t} & , \quad y_\varsigma \neq f_t(x_\varsigma) \end{cases}$$

最终强分类器$F(x) = \sum_t c_t f_t(x)$

6.2.3 GBDT

回归树

训练样本在第i棵树的输入值= 训练样本在第i-1棵树的输入值- 训练样本被第i-1棵树分类的预测值，即每一棵树学的是之前所有树结论和的残差

预测时依次经过所有树

Chapter 7

NLP

7.1 隐含语义分析

7.1.1 LDA

7.1.2 LFM(Latent factor model)

已知$\begin{bmatrix} & \vdots & \\ \cdots & R_{ik} & \cdots \\ & \vdots & \end{bmatrix}$，找出隐含主题分类j

$$\begin{bmatrix} & \vdots & \\ \cdots & R_{ik} & \cdots \\ & \vdots & \end{bmatrix} = \begin{bmatrix} & \vdots & \\ \cdots & P_{ij} & \cdots \\ & \vdots & \end{bmatrix} \begin{bmatrix} & \vdots & \\ \cdots & Q_{jk} & \cdots \\ & \vdots & \end{bmatrix}$$

$$\min_{PQ} \sum_{ik} (R_{ik} - \sum_j P_{ij} Q_{jk})^2 + \lambda_P ||P|| + \lambda_Q ||Q||$$

$$\begin{cases} P_{i'j'} += \eta_P \left(\left(\begin{bmatrix} \cdots & R_{i'k} & \cdots \end{bmatrix} - \begin{bmatrix} \cdots & P_{i'j} & \cdots \end{bmatrix} \begin{bmatrix} & \vdots & \\ \cdots & Q_{jk} & \cdots \\ & \vdots & \end{bmatrix} \right) \begin{bmatrix} \vdots \\ Q^T_{kj'} \\ \vdots \end{bmatrix} - \lambda_P P_{i'j'} \right) \\ Q_{j'k'} += \eta_Q \left(\begin{bmatrix} \cdots & P^T_{j'i} & \cdots \end{bmatrix} \left(\begin{bmatrix} \vdots \\ R_{ik'} \\ \vdots \end{bmatrix} - \begin{bmatrix} & \vdots & \\ \cdots & P_{ij} & \cdots \\ & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ Q_{jk'} \\ \vdots \end{bmatrix} \right) - \lambda_Q Q_{j'k'} \right) \end{cases}$$

7.2 统计语言模型

一段语句的概率

$$p(w_1, \cdots, w_T) = \prod_{t=1}^{T} p(w_t | w_1, \cdots, w_T)$$

或

$$p(w_1, \cdots, w_T) = \prod_{t=1}^{T} p(w_1, \cdots, w_T | w_t)$$

7.2.1 N-gram

假设每个词出现概率仅与它之前的n-1个词有关

$$p(w_t | w_1, \cdots, w_T) \simeq p(w_t | w_{t-n+1}, \cdots, w_{t-1})$$

7.2.2 CBOW

假设每个词出现概率仅与它前后的2n个词有关

$$p(w_t | w_1, \cdots, w_T) \simeq p(w_t | w_{t-n}, \cdots, w_{t+n})$$

7.2.3 Skip-Gram

假设每个词出现概率仅与它前后的2n个词有关

$$p(w_1, \cdots, w_T | w_t) \simeq p(w_t | w_{t-n}, \cdots, w_{t+n})$$

7.2.4 隔词

以上各种模型可不限于紧邻前后，跳过一些词的情况亦可，用于扩展词组和提取远距离信息。可对远距离的词组乘以衰减系数

7.3 词向量

将每个词或者连续几个词表示为坐标空间中的一个点

7.3.1 One-hot Representation

每个词表示为一个向量$(0, \cdots, 0, 1, 0, \cdots, 0)$，向量长度为字典大小。
实践中用Hash表给每个词分配一个编号

7.3.2 Distributed Representation

linear bag-of-words contexts
每个词w表示为一个低维实数向量$\vec{w}$

7.3.2.1 训练

用周围词表示中心词，最大化给定中心词时周围词概率

$$L = \prod_{t=1}^{T} \prod_{-n \leq j \leq n, j \neq 0} p(w_{t+j} | w_t)$$

即

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{-n\le j\le n, j\ne 0}\log p(w_{t+j}|w_t)$$

其中

$$p(w_O|w_I) = \frac{e^{\vec{w}_O\cdot\vec{w}_I}}{\sum_{\vec{w}\in W} e^{\vec{w}\cdot\vec{w}_I}}$$

将$\vec{w}$作为输入，经神经网络输出L。同时训练神经元参数与$\vec{w}$的取值

Chapter 8

其他

8.1 最大熵模型

训练样本$\{(x_\varsigma, y_\varsigma)\}$

已知经验分布$\tilde{P}(x) = \frac{N(x)}{\mathbb{N}}$，$\tilde{P}(x,y) = \frac{N(x,y)}{\mathbb{N}}$。约束条件$I_i(x,y) = \begin{cases} 1 & , \quad \text{x,y满足事实i} \\ 0 & , \quad \text{否则} \end{cases}$。

求贝叶斯分布$P(y|x)$

记后验分布$P(x,y) = P(y|x)\tilde{P}(x)$

$$
\begin{aligned}
& \max_{P(y|x)} H(P(x,y)) = -\sum_{x,y} P(x,y) \log P(x,y) \\
s.t. \quad & E_{P(x,y)}[I_i(x,y)] = E_{\tilde{P}(x,y)}[I_i(x,y)] \\
& \sum_y P(y|x) = 1
\end{aligned}
$$

等价于

$$L(P(x,y),w) = -H(P(x,y)) + w_0(1 - \sum_y P(y|x)) + \sum_i w_i (E_{\tilde{P}(x,y)}[I_i(x,y)] - E_{P(x,y)}[I_i(x,y)])$$

$$\min_{P(y|x)} \max_w L(P(x,y),w)$$

等价于

$$L(P(x,y),w) = -H(P(x,y)) + w_0(1 - \sum_y P(y|x)) + \sum_i w_i (E_{\tilde{P}(x,y)}[I_i(x,y)] - E_{P(x,y)}[I_i(x,y)])$$

$$\max_w \min_{P(y|x)} L(P(x,y),w)$$

等价于（由$\frac{\partial}{\partial P(y|x)} L(P(x,y),w) = 0$得：$P(y|x) = \frac{1}{Z} \exp(\sum_i w_i I_i(x,y))$）

$$
\begin{aligned}
& \max_w L(w) \\
s.t. \quad & I_i(x,y) = 0
\end{aligned}
$$

8.2 评价曲线

二分类问题

预测得到概率大于阈值则划分为1，小于阈值划分为0。每个阈值算出对应概率，为图上一个点。

8.2.0.2 ROC曲线

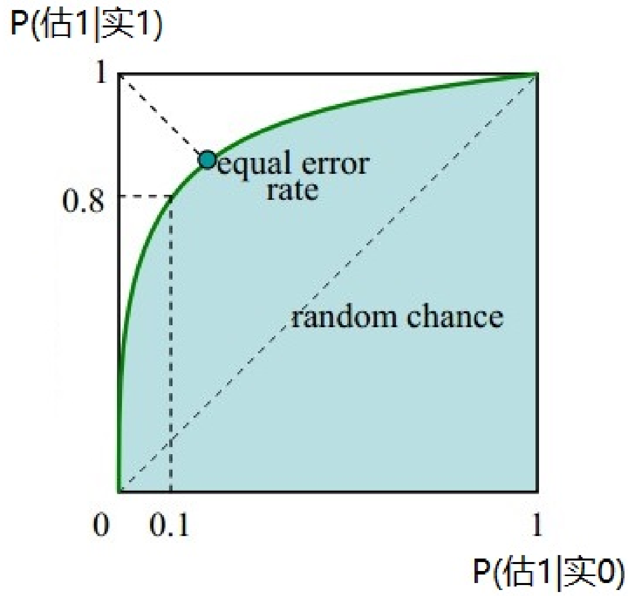

Figure 8.1: ROC

8.2.0.3 PR曲线

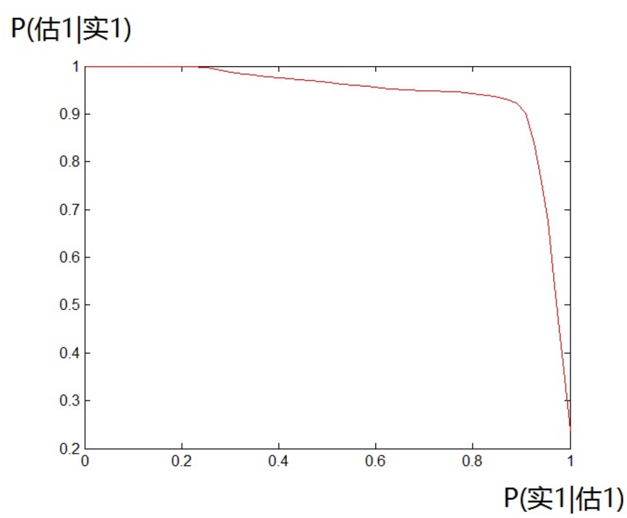

Figure 8.2: PR

Bibliography

[1] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[2] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[4] Hal Daumé III. From zero to reproducing kernel hilbert spaces in twelve pages or less, 2004.