Learning And Mining from DatA
http://lamda.nju.edu.cn

机器学习导论

习题课

詹德川
zhandc@lamda.nju.edu.cn

Outline

- HW5
  - PS1-Bayes Optimal Classier
  - PS2-Naïve Bayes
  - PS3-Bayesian Network
- HW6
  - PS1-Ensemble Methods
  - PS2-Bagging

HW5

PS1-Bayes Optimal Classier

- 试证明在二分类问题中，当两类数据同先验、满足高斯分布且协方差相等时，LDA可产生贝叶斯最优分类器。

Solution. 令$g_i(\mathbf{x}) = \ln(P(c_i)P(\mathbf{x}|c_i))$，其中$y \in \{c_0, c_1\}$，$p(x|c_i) \sim \mathcal{N}(\mu_i, \mathbf{\Sigma})$. 可得，

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}| + \ln P(c_i). \tag{1.1}$$

因此，贝叶斯最优分类器为$f(\mathbf{x}) = g_0(\mathbf{x}) - g_1(\mathbf{x})$，即

$$f(\mathbf{x}) = (\mathbf{\Sigma}^{-1}(\mu_0 - \mu_1))^{\mathrm{T}}\mathbf{x} + b. \tag{1.2}$$

其中，$b = -\frac{1}{2}\mu_0^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mu_0 + \frac{1}{2}\mu_1^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mu_1$. 式1.2与书中3.39一致，证毕.

PS2-Naïve Bayes

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

- 考虑下面的400个训练数据的数据统计情况，其中特征维度为2（$\mathbf{x} = [x_1, x_2]$），每种特征取值0或1，类别标记$y \in \{-1, +1\}$。详细信息如表1所示。

根据该数据统计情况，请分别利用直接查表的方式和朴素贝叶斯分类器给出$\mathbf{x} = [1, 0]$的测试样本的类别预测，并写出具体的推导过程。

表 1: 数据统计信息

x_1	x_2	$y = +1$	$y = -1$
0	0	90	10
0	1	90	10
1	0	51	49
1	1	40	60

PS2-Naïve Bayes

Solution.

(1) 根据表1可知$\mathbf{x} = [1, 0]$, 预测类别为+1.

(2) 首先估计出类先验概率$P(c)$和每个属性的条件概率$P(x_i|c)$:

$$P(y=+1) = \frac{90+90+51+40}{400} \approx 0.678 \ ,$$

$$P(y=-1) = \frac{10+10+49+60}{400} \approx 0.322 \ ,$$

$$P(x_1=1|y=+1) = \frac{51+40}{90+90+51+40} \approx 0.336 \ ,$$

$$P(x_1=0|y=-1) = \frac{49+60}{10+10+49+60} \approx 0.845 \ ,$$

$$P(x_2=0|y=+1) = \frac{90+51}{90+90+51+40} \approx 0.520 \ ,$$

$$P(x_2=0|y=-1) = \frac{10+49}{10+10+49+60} \approx 0.457 \ .$$

PS2-Naïve Bayes

$$P(y=+1)=\frac{90+90+51+40}{400}\approx 0.678\ ,$$

$$P(y=-1)=\frac{10+10+49+60}{400}\approx 0.322\ ,$$

$$P(x_1=1|y=+1)=\frac{51+40}{90+90+51+40}\approx 0.336\ ,$$

$$P(x_1=0|y=-1)=\frac{49+60}{10+10+49+60}\approx 0.845\ ,$$

$$P(x_2=0|y=+1)=\frac{90+51}{90+90+51+40}\approx 0.520\ ,$$

$$P(x_2=0|y=-1)=\frac{10+49}{10+10+49+60}\approx 0.457\ .$$

于是，有

$$P(y=+1)\times P(x_1=1|y=+1)\times P(x_2=0|y=+1)\approx 0.118\ ,$$

$$P(y=-1)\times P(x_1=1|y=-1)\times P(x_2=0|y=-1)\approx 0.124\ .$$

由于$0.118<0.124$, 因此, 朴素贝叶斯分类器将测试样本判别为-1.

PS3-Bayesian Network

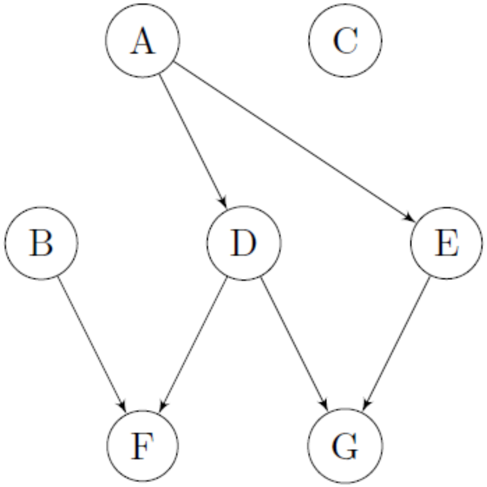
- 贝叶斯网(Bayesian Network)是一种经典的概率图模型，请学习书本7.5节内容回答下面的问题：

  (1) [5pts] 请画出下面的联合概率分布的分解式对应的贝叶斯网结构：

$$\Pr(A,B,C,D,E,F) = \Pr(A)\Pr(B)\Pr(C)\Pr(D|A)\Pr(E|A)\Pr(F|B,D)\Pr(G|D,E)$$

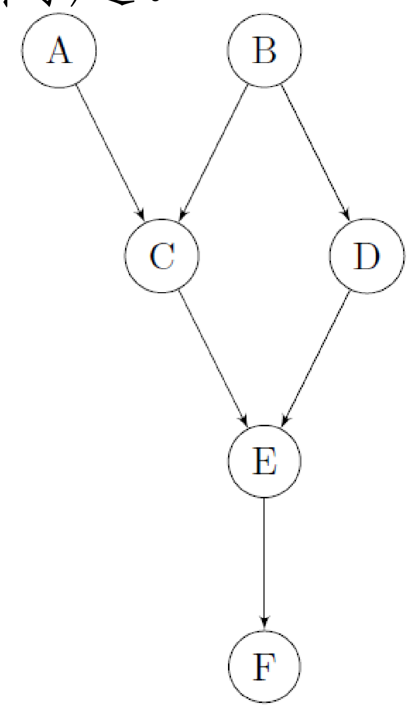
- Solution:

  根据联合概率分布容易画出



PS3-Bayesian Network

- 贝叶斯网(Bayesian Network)是一种经典的概率图模型，请学习书本7.5节内容回答下面的问题：
  (2) [5pts]请写出图1中贝叶斯网结构的联合概率分布的分解表达式。



- Solution:

根据贝叶斯网容易写出联合概率分布的分解表达式如下

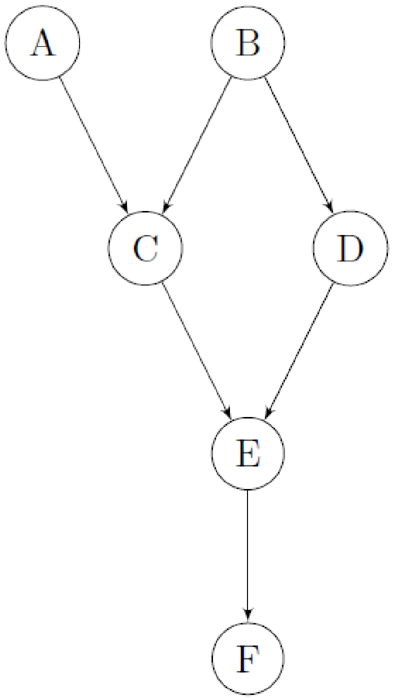
(2) $\Pr(A,B,C,D,E,F) = \Pr(A)\Pr(B)\Pr(C|A,B)\Pr(D|B)\Pr(E|C,D)\Pr(F|E)$

PS3-Bayesian Network

- 贝叶斯网(Bayesian Network)是一种经典的概率图模型，请学习书本7.5节内容回答下面的问题：

  (3) [15pts]基于第(2)问中的图1,

请判断表格2中的论断是否正确.



- Solution:

序号	关系	True/False	序号	关系	True/False
1	$A \perp\!\!\!\perp B$	True	7	$F \perp B \mid C$	False
2	$A \perp B \mid C$	False	8	$F \perp B \mid C, D$	True
3	$C \perp\!\!\!\perp D$	False	9	$F \perp B \mid E$	True
4	$C \perp D \mid E$	False	10	$A \perp\!\!\!\perp F$	False
5	$C \perp D \mid B, F$	False	11	$A \perp F \mid C$	False
6	$F \perp\!\!\!\perp B$	False	12	$A \perp F \mid D$	False

HW6

PS1-Ensemble Methods

- (1) [10pts] 试说明Boosting的核心思想是什么，Boosting中什么操作使得基分类器具备多样性？
- Solution:

通过对基学习器进行迭代训练，在每轮训练中，**通过调整训练样本分布使得分类错误的样本后续受到更多关注**，从而增加基学习器的多样性，最后预测使用所有基学习器加权结合的结果。

使得基学习器具备多样性的操作是权重调整。对样本权重调整后，新一轮的训练相当于基于不同的样本，从而训练出不同于之前的学习器[1]。

1 originally contributed by 刘冰楠

PS1-Ensemble Methods

- (2) [10pts] 试析随机森林为何比决策树Bagging集成的训练速度更快。

- Solution:

随机森林相对于Bagging决策树的关键区别在于，在选择划分属性时，首先随机选择一个属性集的子集，再在这个子集中寻找最优属性。

由于一般随机选择的属性子集规模比所有属性集小(如推荐值$k = \log_2 d$)，训练时只需考察这个较小的子集，从而训练速度更快[1]。

[1] originally contributed by 刘冰楠

PS2-Bagging

考虑一个回归学习任务$f:\mathbb{R}^d \to \mathbb{R}$。假设我们已经学得$M$个学习器$\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \ldots, \hat{f}_M(\mathbf{x})$。我们可以将学习器的预测值看作真实值项加上误差项

$$\hat{f}_m(\mathbf{x}) = f(\mathbf{x}) + \epsilon_m(\mathbf{x}) \tag{2.1}$$

每个学习器的期望平方误差为$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]$。所有的学习器的期望平方误差的平均值为

$$E_{av} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] \tag{2.2}$$

M个学习器得到的Bagging模型为

$$\hat{f}_{bag}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \hat{f}_m(\mathbf{x}) \tag{2.3}$$

Bagging模型的误差为 $$\epsilon_{bag}(\mathbf{x}) = \hat{f}_{bag}(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x}) \tag{2.4}$$

其期望平均误差为 $$E_{bag} = \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2] \tag{2.5}$$

PS2-Bagging

Learning And Mining from DatA
http://lamda.nju.edu.cn

(1) **[10pts]** 假设$\forall\, m \neq l$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0$，$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$。证明

$$E_{bag} = \frac{1}{M} E_{av} \tag{2.6}$$

Proof.

(1) $E_{bag} = \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2] = \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}[\sum_{i=1}^{M}\sum_{j=1}^{M}\epsilon_i(\mathbf{x})\epsilon_j(\mathbf{x})] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] = E_{av}.$

PS2-Bagging

(2) **[10pts]** 试证明不需对$\epsilon_m(\mathbf{x})$做任何假设，$E_{bag} \leq E_{av}$始终成立。（提示：使用Jensen's inequality）

Proof.

(2) 由Jensen's inequality可知，对于凸函数φ，有

$$\varphi(\sum_{m=1}^{M} a_m x_m) \leq \sum_{m=1}^{M} a_m \varphi(x_m) \tag{2.7}$$

其中$a_i \geq 0, \sum_{m=1}^{M} a_i = 1$。文中所用函数$\varphi(x) = x^2$是凸函数，因此

$$E_{bag} = \mathbb{E}_{\mathbf{x}}[(\frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x}))^2] \leq \mathbb{E}_{\mathbf{x}}[\frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x})^2] = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2] = E_{av}.$$

□

Q & A

Thanks !