习题一

141242006，袁帅

2017 年 3 月 15 日

Problem 1

若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设，此时的版本空间是什么？在此情形下，试设计一种归纳偏好用于假设选择。

Solution. Given the noise in training set, if no hypothesis is consistent with the training samples, the version space would be an empty set by definition.

In regard to the selection of inductive bias, according to *No Free Lunch Theorem*, there's actually no significant differences among these biases, for the absence of specific problem settings. In this case, generally speaking, an intuitive method would be to examine the extent to which a given hypothesis h is shifted from the ground-truth f. Specifically, an error function $err_{h,f}(\boldsymbol{x})$ should be defined, for all $\boldsymbol{x} \in \mathcal{X}$, and we will find the h^* with the minimum sum of error on the training set X, i.e.

$$h^* = \arg\min_h \sum_{\boldsymbol{x}\in\mathcal{X}} err_{h,f}(\boldsymbol{x}), \tag{1}$$

in which $err_{h,f}(\boldsymbol{x})$ could be defined as follow:

- In classification problem, we can simply count the number of errors made by a specific hypothesis. That gives $err_{h,f}(\boldsymbol{x}) = \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$.
- In regression problem, we can apply least square method, i.e. $err_{h,f}(\boldsymbol{x}) = (h(\boldsymbol{x}) - f(\boldsymbol{x}))^2$.

A possible pitfall in this method is that the same minimum sum of error may be reached by more than one hypothesis, so when dealing with such ties, we need another inductive bias to decide. In this case, *Occam's razor* would do the job, or we can also randomly choose one in all h^* candidates.

Problem 2

对于有限样例，请证明

$$\mathrm{AUC} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right)$$

Proof. Without loss of generality, we assume that all samples $x_k \in D$ are sorted by their prediction value, i.e. $f(x_1) \geq f(x_2) \geq \cdots \geq f(x_m)$. In this notion, the definition of AUC [1] could be rewritten as

$$\mathrm{AUC} = \frac{1}{2} \sum_{k=1}^{m-1} (\mathrm{FPR}_{k+1} - \mathrm{FPR}_k) \cdot (\mathrm{TPR}_k + \mathrm{TPR}_{k+1}), \tag{2}$$

where FPR_k and TPR_k are false positive rate and true positive rate respectively, after we introduce $\boldsymbol{s}_k$ into the plot of ROC curve.

By definition, we have $TP + FN = m^+$ and $TN + FP = m^-$, which yield

$$\mathrm{TPR}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} = \frac{1}{m^+} \sum_{x^+ \in D^+} \mathbb{I}(f(x^+) \geq f(x_k)), \tag{3}$$

$$\mathrm{FPR}_k = \frac{\mathrm{FP}_k}{\mathrm{TN}_k + \mathrm{FP}_k} = \frac{1}{m^-} \sum_{x^- \in D^-} \mathbb{I}(f(x^-) \geq f(x_k)). \tag{4}$$

Therefore, we could obtain

$$\begin{aligned} \mathrm{TPR}_k + \mathrm{TPR}_{k+1} &= \frac{1}{m^+} \sum_{x^+ \in D^+} \left(\mathbb{I}(f(x^+) \geq f(x_k)) + \mathbb{I}(f(x^+) \geq f(x_{k+1})) \right) \\ &= \frac{1}{m^+} \sum_{x^+ \in D^+} \left(2\mathbb{I}(f(x^+) \geq f(x_k)) + \mathbb{I}(f(x_k) > f(x^+) \geq f(x_{k+1})) \right) \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{FPR}_{k+1} - \mathrm{FPR}_k &= \frac{1}{m^-} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^-) \geq f(x_{k+1})) - \mathbb{I}(f(x^-) \geq f(x_k)) \right) \\ &= \frac{1}{m^-} \sum_{x^- \in D^-} \mathbb{I}(f(x_k) > f(x^-) \geq f(x_{k+1})). \end{aligned} \tag{6}$$

[1] Eq.(2.20) on textbook, page 35.

Plug Eq.(5) and Eq.(6) into Eq.(2), we have

$$\begin{aligned}
\text{AUC} &= \frac{1}{2m^+m^-}\sum_{k=1}^{m-1}\sum_{x^+\in D^+}\big(2\mathbb{I}(f(x^+)\geq f(x_k))+\mathbb{I}(f(x_k)>f(x^+)\geq f(x_{k+1}))\big)\\
&\quad\cdot\sum_{x^-\in D^-}\mathbb{I}(f(x_k)>f(x^-)\geq f(x_{k+1}))\\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\sum_{k=1}^{m-1}[(\mathbb{I}(f(x^+)\geq f(x_k))+\frac{1}{2}\mathbb{I}(f(x_k)>f(x^+)\geq f(x_{k+1})))\\
&\quad\cdot\mathbb{I}(f(x_k)>f(x^-)\geq f(x_{k+1}))]\\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\sum_{\substack{1\leq k\leq m-1\\ f(x_k)\neq f(x_{k+1})}}[(\mathbb{I}(f(x^+)\geq f(x_k))+\frac{1}{2}\mathbb{I}(f(x_k)>f(x^+)\geq f(x_{k+1})))\\
&\quad\cdot\mathbb{I}(f(x_k)>f(x^-)\geq f(x_{k+1}))]\\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\sum_{\substack{1\leq k\leq m-1\\ f(x_k)\neq f(x_{k+1})}}[(\mathbb{I}(f(x^+)>f(x_{k+1}))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x_{k+1})))\\
&\quad\cdot\mathbb{I}(f(x^-)=f(x_{k+1}))]\\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\sum_{\substack{1\leq k\leq m-1\\ f(x_k)\neq f(x_{k+1})\\ f(x^-)=f(x_{k+1})}}[\mathbb{I}(f(x^+)>f(x_{k+1}))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x_{k+1}))]\\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}[(\mathbb{I}(f(x^+)>f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))]. \qquad (7)
\end{aligned}$$

Note that in the derivation of Eq.(7), we used a trick that the assertions in the indicator functions $\mathbb{I}(\cdot)$ could sometimes be rewritten into the summation's index restriction, since if $\mathbb{I}(\cdot)=0$, the term will contribute 0 to the summation.[2]

□

Problem 3

在某个西瓜分类任务的验证集中，共有 10 个示例，其中有 3 个类别标记为“1”，表示该示例是好瓜；有 7 个类别标记为“0”，表示该示例不是好瓜。由于学习方法能力有限，我们只能产生在验证集上精度 (accuracy) 为 0.8 的分类器。

(a) 如果想要在验证集上得到最佳查准率 (precision)，该分类器应该作出何种预测?

此时的查全率 (recall) 和 F1 分别是多少?

(b) 如果想要在验证集上得到最佳查全率 (recall)，该分类器应该作出何种预测?

此时的查准率 (precision) 和 F1 分别是多少?

Solution. (a) By definition, the number of positive samples is $TP+FN=3$, and that of negative samples is $TN+FP=7$. The accuracy is $\frac{TP+TN}{TP+TN+FP+FN}=0.8$, i.e.

[2]For instance, $\sum_{k=1}^{n} g(k)\cdot\mathbb{I}(k \text{ is odd})=\sum_{1\leq k\leq n,k \text{ is odd}} g(k)$, since if k is even, that will contribute $g(k)\cdot 0=0$ to the summation.

$TP+TN=8$. We can immediately obtain $TN=8-TP$, $FP=TP-1$, $FN=3-TP$ and the precision

$$P=\frac{TP}{TP+FP}=\frac{TP}{2TP-1}=\frac{1}{2-\frac{1}{TP}}\leq 1, \tag{8}$$

in which the equality holds when $TP=1$.

Therefore, the model should return positive for only the most positive-likely sample and negative for the other samples. In this case, the recall $R=\frac{TP}{TP+FN}=\frac{1}{3}$, and $F1=\frac{2P\cdot R}{P+R}=0.5$.

(b) The recall

$$R=\frac{TP}{TP+FN}=\frac{TP}{3}\leq 1, \tag{9}$$

in which the equatlity holds when $TP=3$.

Therefore, the model should return positive for the 5 most positive-likely samples and negative for the other samples. In this case, the precision $P=\frac{TP}{TP+FP}=0.6$, and $F1=\frac{2P\cdot R}{P+R}=0.75$.

Problem 4

在数据集 D_1, D_2, D_3, D_4, D_5 运行了 A, B, C, D, E 五种算法，算法比较序值表如表 1所示:

表 1: 算法比较序值表

数据集	算法 A	算法 B	算法 C	算法 D	算法 E
D_1	2	3	1	5	4
D_2	5	4	2	3	1
D_3	4	5	1	2	3
D_4	2	3	1	5	4
D_5	3	4	1	5	2
平均序值	3.2	3.8	1.2	4	2.8

使用 Friedman 检验 ($\alpha=0.05$) 判断这些算法是否性能都相同。若不相同，进行 Nemenyi 后续检验 ($\alpha=0.05$)，并说明性能最好的算法与哪些算法有显著差别。

Solution. By Friedman Test($\alpha=0.05$), substituting $N=5$, $k=5$, $\boldsymbol{r}=(3.2, 3.8, 1.2, 4, 2.8)'$, we have $\tau_{\chi^2}=\frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2-\frac{k(k+1)^2}{4}\right)=9.92$ and $\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}=3.9365$. Since $\tau_F=3.9365>F_{0.05}(4,16)=3.0069$, we reject the null hypothesis H_0 and claim that these five algorithms are significantly different in performance.

By Nemenyi Post-hoc Test($\alpha = 0.05$), according to Tab.(2.7) on textbook, page 43, we get $q_\alpha = 2.728$, and $CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} = 2.728$. The best algorithm is Algorithm C, since it has the best average rank. Consequently, because $|4 - 1.2| > 2.728$, Algorithm C has significant differences from Algorithm D.