Week 1 Report

Li Zhengyuan
Tsinghua University
zhengyua17@mails.tsinghua.edu.cn

Abstract

Notes for an intro to ML.

1 Basic concepts

1.1 Definition

To build a system that reliably improves its performance P at task T, following experience E.

1.2 Classification by whether the data has labels

- Supervised learning (classification & regression)
- Semi-supervised learning
- Unsupervised learning(clustering)

New preprint option for 2018

2 Evaluation Methods

2.1 Hold-out

- Dataset = training set (S)+ testing set (T)
- But a dilemma is what proportion of training set we should choose.

2.2 Cross-validation

- Divide the dataset into k subsets(mutually exclusive)
- Each time, the combination of (k - 1) subsets is used as the training set, the last one as the testing set
- a.k.a K-fold cross validation
- More advanced than hold-out

2.3 Bootstrapping

- Dataset D (m samples) -> generate D'
- For i in range (m)
  Randomly select a sample s, and place its copy into D'
  End
- About 36.8% of samples don't appear in D'.
- May choose one sample twice.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})+\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \\
&\quad +\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y+y-y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)^2\right]+\mathbb{E}_D\left[\left(y-y_D\right)^2\right] \\
&\quad +2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)\left(y-y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\left(\bar{f}(\boldsymbol{x})-y\right)^2+\mathbb{E}_D\left[\left(y_D-y\right)^2\right]\ ,
\end{aligned}
$$

Figure 1: Decomposition

3 Bias-variance decomposition

3.1 Notations

$y_D := labelinthedataset$

$y := thereallabel$

$\bar{f}(x) = E_D f[x;D]$

$var(x) = E_D[(f(x;D) - \bar{f}(x))^2]$

$\epsilon^2 = E_D[(y_D - y)^2]$

$bias^2(x) = (\bar{f}(x) - y)^2$

3.2 Decomposition

To see the figure 1.

3.3 Explanation

- Noise, which depends on the problem itself(e.g. P(xly) is stochastic.), gives a lower bound of generalization error.
- Bias describes the ability of the model.
- Variance describes the perturbation of data.
- When the model underfits, bias is the main cause of generalization error;when it overfits, variance is of importance.

4 Decision Tree

4.1 Information Gain

- We use **Entropy** to measure the purity of the sample,which is

$$Ent(D) = -\sum_{k=1}^{|y|} p_k * log_2 p_k \tag{1}$$

输入: 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
属性集 $A = \{a_1, a_2, \ldots, a_d\}$.

过程: 函数 TreeGenerate(D, A)

生成结点 node;
if D 中样本全属于同一类别 C **then**
  将 node 标记为 C 类叶结点; **return**
end if
if $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
  将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; **return**
end if
从 A 中选择最优划分属性 a_*;
for a_* 的每一个值 a_*^v **do**
  为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
  if D_v 为空 **then**
    将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; **return**
  else
    以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
  end if
end for

输出: 以 node 为根结点的一棵决策树

Figure 2: Decision Tree

- So when we classify the sample by attributes, we get more information about them.

4.2 Algorithm

To see the figure 2.

4.3 Regularization

Prepruning To estimate before dividing : if dividing cannot improve the ability, mark the current node as a leaf node.

Postpruning To buttom-up analyze those non-leaf nodes : if replacing the node with a leaf node can improve the ability, then do it.