SNS COLLEGE OF TECHNOLOGY

Coimbatore-37.
An Autonomous Institution

COURSE NAME : 19CST301 & INTRODUCTION TO MACHINE LEARNING

III YEAR/ V SEMESTER

UNIT – 3 DEEP LEARNING

Topic: Interpretation

Mrs.S.R.Janani
Assistant Professor
Department of Computer Science and Engineering

Interpretation

- Single decision trees are highly interpretable.
- The entire model can be completely represented by a simple two-dimensional graphic (binary tree) that is easily visualized.
- Linear combinations of trees lose this important feature, and must therefore be interpreted in a different way.
  - Relative Importance of Predictor Variables
  - Partial Dependence Plots

Relative Importance of Predictor Variables

- In data mining applications the **input predictor variables** are seldom equally relevant.
- Often only a few of them have substantial influence on the response; the **vast majority are irrelevant** and could just as well have not been included.
- It is often useful **to learn** the relative importance or contribution of each input variable in predicting the response

For a single decision tree T, Breiman et al. (1984) proposed

$$\mathcal{I}_\ell^2(T) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 \, I(v(t) = \ell) \qquad (10.42)$$

as a measure of relevance for each predictor variable Xℓ.

The sum is over the J − 1 internal nodes of the tree.

At each such node t, one of the input variables Xv(t) is used to partition the region associated with that node into two subregions; within each a separate constant is fit to the response values.

The particular variable chosen is the one that gives maximal estimated improvement $\hat{\imath}_t^2$ in **squared error risk** over that for a constant fit over the entire region.

The squared relative importance of variable Xℓ is the sum of such squared improvements over all internal nodes for which it was chosen as the splitting variable.

This importance measure is easily generalized to additive tree expansions. it is simply averaged over the trees

$$\mathcal{I}_\ell^2 = \frac{1}{M}\sum_{m=1}^{M}\mathcal{I}_\ell^2(T_m). \qquad (10.43)$$

Due to the **stabilizing effect of averaging,** this measure turns out to be more reliable than is its counterpart for a single tree.

Also, because of shrinkage the masking of important variables by others with which they are **highly correlated is much less of a problem.**

Note that (10.42) and (10.43) refer to squared relevance; the actual relevances are their respective square roots.

Since these measures are relative, it is customary to assign the largest a value of 100 and then scale the others accordingly.

Figure 10.6 shows the relevant importance of the 57 inputs in predicting spam versus email.

- For K-class classification, K separate models $f_k(x)$,k = 1, 2,... ,K are induced, each consisting of a sum of trees

$$f_k(x) = \sum_{m=1}^{M} T_{km}(x). \tag{10.44}$$

In this case (10.43) generalizes to

$$\mathcal{I}_{\ell k}^2 = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_\ell^2(T_{km}). \tag{10.45}$$

Here $\mathcal{I}_{\ell k}$ is the relevance of X_ℓ in separating the class k observations from the other classes. The overall relevance of X_ℓ is obtained by averaging over all of the classes

$$\mathcal{I}_\ell^2 = \frac{1}{K} \sum_{k=1}^{K} \mathcal{I}_{\ell k}^2. \tag{10.46}$$

Partial Dependence Plots

- After the most relevant variables have been identified, the next step is to attempt to understand the nature of the dependence of the approximation f(X) on their joint values.
- Graphical renderings of the f(X) as a function of its arguments provides a comprehensive summary of its dependence on the joint values of the input variables.
- Unfortunately, such visualization is limited to low-dimensional views.
- We can easily display functions of one or two arguments, either continuous or discrete (or mixed), in a variety of different ways; this book is filled with such displays.
- Functions of slightly higher dimensions can be plotted by conditioning on particular sets of values of all but one or two of the arguments, producing a trellis of plots (Becker et al., 1996).1

- For more than two or three variables, viewing functions of the corresponding higher-dimensional arguments is more difficult.
- A useful alternative can sometimes be to view a collection of plots, each one of which shows the partial dependence of the approximation f(X) on a selected small subset of the input variables.
- Although such a collection can seldom provide a comprehensive depiction of the approximation, it can often produce helpful clues, especially when f(x) is dominated by low-order interactions

Consider the subvector XS of ℓ < p of the input predictor variables

XT = (X1,X2,... ,Xp), indexed by S ⊂ {1, 2,... ,p}.

Let C be the complement set, with S ∪ C = {1, 2,... ,p}.

A general function f(X) will in principle depend on all of the input variables: f(X) = f(XS ,XC).

One way to define the average or partial dependence of f(X) on XS is

$$f_S(X_S) = \mathrm{E}_{X_C} f(X_S, X_C). \qquad (10.47)$$

- This is a marginal average of f, and can serve as a useful description of the effect of the chosen subset on f(X) when,
- for example, the variables in XS do not have strong interactions with those in XC.
- Partial dependence functions can be used to interpret the results of any “black box” learning method.
- They can be estimated by

$$\bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{iC}), \qquad (10.48)$$

- where {x1C,x2C,... ,xNC} are the values of XC occurring in the training data. This requires a pass over the data for each set of joint values of XS for which ¯fS (XS) is to be evaluated.

- This can be computationally intensive, even for moderately sized data sets. Fortunately with decision trees, ¯fS (XS) (10.48) can be rapidly computed from the tree itself without reference to the data .
- It is important to note that partial dependence functions defined in (10.47) represent the effect of XS on f(X) after accounting for the (average) effects of the other variables XC on f(X).
- They are not the effect of XS on f(X) ignoring the effects of XC. The latter is given by the conditional expectation

$$\tilde{f}_S(X_S) = \mathrm{E}(f(X_S, X_C)|X_S), \qquad (10.49)$$

- and is the best least squares approximation to f(X) by a function of XS alone.
- The quantities ˜fS (XS) and ¯fS (XS) will be the same only in the unlikely event that XS and XC are independent.
- For example, if the effect of the chosen variable subset happens to be purely additive,

$$f(X) = h_1(X_S) + h_2(X_C). \qquad (10.50)$$

- Then (10.47) produces the h1(XS) up to an additive constant. If the effect is purely multiplicative,

$$f(X) = h_1(X_S) \cdot h_2(X_C), \qquad (10.51)$$

- then (10.47) produces h1(XS) up to a multiplicative constant factor.
- On the other hand, (10.49) will not produce h1(XS) in either case.
- In fact, (10.49) can produce strong effects on variable subsets for which f(X) has no dependence at all.

- Viewing plots of the partial dependence of the boosted-tree approximation (10.28) on selected variables subsets can help to provide a qualitative description of its properties.
- Illustrations are shown in Sections 10.8 and 10.14. Owing to the limitations of computer graphics, and human perception, the size of the subsets XS must be small (l ≈ 1, 2, 3).
- There are of course a large number of such subsets, but only those chosen from among the usually much smaller set of highly relevant predictors are likely to be informative.
- Also, those subsets whose effect on f(X) is approximately additive (10.50) or multiplicative (10.51) will be most revealing.

- For K-class classification, there are K separate models (10.44), one for each class. Each one is related to the respective probabilities (10.21) through

$$f_k(X) = \log p_k(X) - \frac{1}{K}\sum_{l=1}^{K} \log p_l(X). \qquad (10.52)$$

- Thus each fk(X) is a monotone increasing function of its respective probability on a logarithmic scale.
- Partial dependence plots of each respective fk(X) (10.44) on its most relevant predictors (10.45) can help reveal how the log-odds of realizing that class depend on the respective input variables.

References

- AlpaydinEthem, “Introduction to Machine Learning”, MIT Press, Second Edition, 2010.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer; Second Edition, 2009.


Thank
you