目录

Part I

Deep Learning

Chapter 1

Natural language processing, NLP

1.1 tf-idf

在一份给定的文件里，词频（term frequency, tf）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数（term count）的标准化，以防止它偏向长的文件。（同一个词语在长文件里可能会比段文件有更高的次数，而不管该词语重要与否）对于在某一个特定文件里的词语 t_i 来说，它的重要性可表示：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1.1}$$

式中假设文件 d_j 中共有 k 个词语， $n_{k,j}$ 是 t_k 在文件 d_j 中出现的次数。分子 $n_{i,j}$ 是该词在文件 d_j 中出现次数，而分母则是在文件 d_j 中所有字词出现的次数之和。

逆向文件频率（inverse document frequency, idf）是一个词语普遍重要性的度量。某一特定词语的 idf，可以有总文件数目除以包含该词语之文件的数目，再将得到的上取以10为底的对数得到：

$$idf_i = \log_{10} \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{1.2}$$

式中，$|D|$表示语料库中的文件总数，$|\{j : t_i \in d_j\}|$ 表示包含词语 t_i 的文件数目，如果词语不再资料中，就导致分母为零，因此一般情况下使用$1 +$

$|\{j : t_i \in d_j\}|$，然后

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{1.3}$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 $tf - idf$，因此，$tf - idf$ 倾向于过滤掉常见的词语，保留重要的词语。

例子：假如一篇文件的总词数是100个，而词语母牛出现了3次，那么母牛一次在该文件中的词频就 $3/100 = 0.03$，而计算文件频率的方法就是以文件集的文件总数除以出现母牛一词的文件数，所以，如果母牛一次在1000份文件出现过，而文件总数是10000000份的话，其你想文件频率就是$\log_{10}(10000000/1000) = 4$，最后的 $tf - idf$的分数为 $0.03 * 4 = .012$，更多访问 tf-idf。

Part II

Machine Learning

Chapter 2

决策树

2.1 The metrics for measuring a split

2.1.1 Gini Impurity

Definition 1 (Gini不纯度) *Gini Impurity, as its name implies, measures the impurity rate of the class distribution of data points, or the class mixture rate. For a dataset with K classes, suppose that data from class $k (1 \leq k \leq K)$ takes up a fraction $f_k (0 \leq f_k \leq 1)$ of the entire dataset; then the Gini Impurity of this dataset is written as follows:*

$$\textit{Gini impurity} = 1 - \sum_{k=1}^{K} f_k^2$$

A lower Gini Impurity indicates a purer dataset.

2.1.2 Information Gain

Definition 2 *Entropy is a probabilistic measure of uncertainty. Given a K-class dataset, and $f_k (0 \leq f_k \leq 1)$ denoted as the fraction of data from class $k (1 \leq k \leq K)$, the entropy of the dataset is defined as follows:*

$$\textit{Entropy} = -\sum_{k=1}^{K} f_k * \log_2 f_k$$

Definition 3 (Information Gain) *Information Gain, measures the improvement of purity after splitting or, in other words, the reduction of uncertainty due to a split. Higher Information Gain implies better splitting. We obtain the Information Gain of a split by comparing the entropy before and after the split.*

$$\begin{aligned} Information\,Gain &= Entropy(before) - Entropy(After) \\ &= Entropy(Parent) - Entropy(Children) \end{aligned}$$

Lower entropy implies a purer dataset with less ambiguity.

Chapter 3

Logistic regression

3.1 odds

Let's first introduce the odds: the odds in favor of a particular event. The odds can be written as$\frac{p}{1-p}$, where p stands for the probability of the positive event. The term “positive event” does not necessarily mean “good,” but refers to the event that we want to predict.

Chapter 4

Methods

4.1 MultiClass

4.1.1 The OvA method for multi-class classification

OvA, which is sometimes also called **one-versus-rest (OvR)**, is a technique that allows us to extend any binary classifier to multi-class problems. Using OvA, we can train one classifier per class, where the particular class is treated as the positive class and the examples from all other classes are considered negative classes. If we were to classify a new, unlabeled data instance, we would use our n classifiers, where n is the number of class labels, and assign the class label with the highest confidence to the particular instance we want to classify.

Chapter 5

Introduction

5.1 Machine Learning Category

5.1.1 Parametric versus non-parametric models

Machine learning algorithms can be grouped into parametric and non-parametric models. Using parametric models, we estimate parameters from the training dataset to learn a function that can classify new data points without requiring the original training dataset anymore. Typical examples of parametric models are the perceptron, logistic regression, and the linear SVM. In contrast, non-parametric models can't be characterized by a fixed set of parameters, and the number of parameters changes with the amount of training data. Two examples of non-parametric models that we have seen so far are the decision tree classifier/random forest and the kernel (but not linear) SVM.

KNN belongs to a subcategory of non-parametric models described as instance-based learning. Models based on instance-based learning are characterized by memorizing the training dataset, and lazy learning is a special case of instance-based learning that is associated with no (zero) cost during the learning process.

Chapter 6

Metric

6.1 Distance

6.1.1 minkowski

The minkowski distance is just a generalization of the Euclidean and Manhattan distance, which can be written as follows:

$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sqrt[p]{\sum_k |x_k^{(i)} - x_k^{(j)}|^p} \tag{6.1}$$

It becomes the Euclidean distance if we set the parameter $p = 2$ or the Manhattan distance at $p = 1$.