Machine learning exam 2017

Henrik Lund Mortensen

December 29, 2017

1 Linear Models

2 Learning Theory

- Supervised learning
- In-sample error vs out-of-sample error

Learning feasibility

In general, we have an unknown function $f(x)$ and only a finte set of samples $\mathcal{D} = \{x_i, f(x_i)\}$. Our job is to use $\mathcal{D}$ to estimate f outisde of $\mathcal{D}$. Consider a set of hypotheses, $\mathcal{H}$. The goal is to choose the hypothesis, $g \in \mathcal{H}$ that approximates $f(x)$ best. There are two thing we must consider:

1. Can we find a g such that the training samples are well explained, i.e. low $E_{\text{in}}(g)$?
2. Can we make sure that the $E_{\text{out}}(g)$ is close to $E_{\text{in}}(g)$?

The answer to the first question depends on the complexity of f and the $\mathcal{H}$. It is clear that we can always find a set $\mathcal{H}$ which contains a hypothesis that matches the training data perfectly, i.e. $E_{\text{in}} = 0$. However, we get in trouble in the part 2 above. We cannot guarantee that $E_{\text{out}}(g)$ is close to $E_{\text{in}}(g)$, but we can estimate the probability that is it. This is given by the Hoeffding bound

$$\mathcal{P}\left[|E_{\text{out}}(g) - E_{\text{in}}(g)| > \epsilon\right] \leq 2Me^{-2\epsilon^2 N}, \tag{1}$$

where N is the number of samples in $\mathcal{D}$ and M is the size of $\mathcal{H}$ (number of hypotheses). There is a trade-off between choosing $\mathcal{H}$ complex enough to describe f well, while low enough to have a reasonable bound in Eq. (1). The Hoeffding bound only applies to finte $\mathcal{H}$, as it becomes meaningless when $M \to \infty$.

When the hypothesis space becomes infinite we must replace M by something else. We call the replacement the *growth function*. We restrict outselves to binary target functions, such that the target function (and our hypotheses, $h \in \mathcal{H}$) map from $\mathcal{X}$ to $\{+1, -1\}$. Further, we define a dichotomy as an N-tuple of $+1$'s and -1's such that the dichotomies generated by $\mathcal{H}$ on some set $x_1, x_2, \ldots, x_N \in \mathcal{X}$ is given by

$$\{h(x_1), h(x_2), \ldots, h(x_N) | \forall h \in \mathcal{H}\}. \tag{2}$$

It is clear that there can be a maximum of 2^N different dichotomies for N points. However, it is not true that any $\mathcal{H}$ can generate all dichotomies. We say that if $\mathcal{H}$ generates all possible dichotomies for some set, it *shatters* this set.

3 Support Vector Machines

In a binary classifier labels inputs either $+1$ or -1 (for example). However, points close to the decision boundary are more uncertain that those far from it.

4 Neural Nets

5 Decision Trees and Ensemble Methods

6 Hidden Markov Models - Decoding

7 Hidden Markov Models - Training

8 Unsupervised Learning - Clustering

9 Unsupervised Learning - Outlier Detection and Dimensionality Reduction