50.021 – AI

Alex

Week 03: Overfitting and convolutions

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity, every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

Due: week4Thursday, 13th of June, 6pm

1 A quick look on VC-dimension (easy!)

VC-Dimension was historically one of the first complexity measures for classifiers.

We say that a function class of classifiers $\mathcal{F}$ has VC dimension K if there exists one set of K data points $(x_1, \ldots, x_K)$ such that for every possible labelling $y_1, \ldots, y_K$ of these K data points, there exists a classifier $f \in \mathcal{F}$ which achieves zero training error on this one set.

Question 1 to you: What is the VC-Dimension of the set $\mathcal{F}$ given by the set of all decision trees with 4^R leaves and such that each leaf can have labels $y \in \{-1, +1\}$?

Question 2 to you: What is the VC-Dimension of the set $\mathcal{F}$ given by the set of all decision trees with 4^R leaves and such that each second leaf can have labels $y \in \{-1, +1\}$, while the other half have labels $y = -1$?

Historical note, (just for information): the VC dimension V was the first complexity measure which allowed to establish bounds on the expected test error $E_{(x,y)\sim P_{test}}[I[f(x) \neq y]]$ in the form of statements like:

For any δ with probability $1-\delta$ over draws of training data sets of size n it holds that

$$E_{(x,y)\sim P_{test}}[I[f(x)\neq y]] \leq \frac{1}{n}\sum_{i=1}^{n} I[f(x_i)\neq y_i] + \sqrt{\frac{\log(n)}{n}(C_1 V - C_2\log(\delta))}$$

That inequality states, that the expected test error $E_{(x,y)\sim P_{test}}[I[f(x)\neq y]]$ under the unknown P_{test} can be bounded by the average error on the training data set plus some terms depending on the VC-dimension and the "statement does not hold"-probability δ (and the sample size n). For newer results you can consider e.g. the Rademacher complexity.

Question 3 to you: What is the limit of $\frac{\log(n)}{n}$ as $n\to\infty$? Show some work (e.g. l'hospital rule) to compute it.

2 Overfitting with more and more dimensions

Lets consider the case when we have a fixed number of datapoints n and we go into more and more high dimensional spaces.
More precisely:

- we have a classification problem with samples (x,y) with $y\in\{-1,+1\}$ being the labels.
- Suppose for now that we have a one-dimensional feature $x_i=(x_i^{(1)})$ where $x^{(1)}$ denotes the index for the only dimension, and the subscript $_i$ in x_i is the number of the sample. I introduce this notation, because we will consider soon samples in D dimensions $x_i=(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(D)})$. Consider the following distribution of samples.

$$P(X^{(1)}<0|Y=-1)=0.5$$
$$P(X^{(1)}<0|Y=+1)=0.5$$

  This tells that the classifier

$$f_0(x)=2I[x^{(1)}\geq 0]-1=\begin{cases}-1 & x^{(1)}<0\\ +1 & x^{(1)}\geq 0\end{cases}$$

  is not that excessively useful as a predictor under the expectation under $P(x,y)$.
- compute $E_{(x,y)\sim P}[I[f_0(x)\neq y]]$. Show your work in detail. This works for any value of $P(Y=+1)$.
- Suppose we draw the N samples statistically independently. Let the first $N/2$ points be of class -1.

What is the probability that we draw N samples such that the error on this training dataset is zero under $f_0(x)$? Express this event in terms of conditions to x_i for the first $N/2$ points and for the last $N/2$ points. Then compute its probability under above $P(X|Y)$.

- now lets consider a D-dimensional setup. $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)})$

$$P(X^{(d)} < 0|Y = -1) = 0.5 \; \forall d = 1, \ldots, D$$
$$P(X^{(d)} < 0|Y = +1) = 0.5 \; \forall d = 1, \ldots, D$$

and all the dimensions are statistically independent, thus e.g.

$$P(X^{(d_1)} < 0, X^{(d_2)} < 0, X^{(d_3)} < 0|Y) = \prod_{k=1}^{3} P(X^{(d_k)} < 0|Y)$$

From the $D = 1$ case above you know the distribution of the case when in one of these D dimensions the error on this training dataset is zero under $f_0(x^d)$.

- What is the probability distribution that we draw N samples such that in exactly K out of D dimensions (remember $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)})$) $\{d_1, \ldots, d_K\} \subset 1, \ldots, D$ $f_0(x^{(d_k)})$ achieves zero training error? Give its name and its parameters.
- What is the precise probability that we draw N samples such that in at least one dimension d out of D dimensions $f_0(x^{(d)})$ achieves zero training error?
- What is the limit of this probability as $D \to \infty$? What is the $\mathcal{O}(\cdot)$ complexity of the convergence of this limit as a function of D ?

Hope that tells you something about spurious correlations in high dimensions.

3 Convolutions (yaaawn)

- Suppose your input feature map has 52 channels with height 228 and width 137. Suppose you use on it a 2d convolution with kernel height 17 and kernel width 15 with strides 5 (height) and 3 (width), and padding of 3 for both, and 15 output channels. What is the output feature map size in terms of (ch, h, w)?

- now you use on top of that output of the 2d-conv a pooling with kernel size $3, 3$ and stride 2. What is the output feature map size in terms of (ch, h, w)?

- What is the number of parameters in the convolution layer when biases are used?