Introduction to Machine Learning

Homework 1

Dec 7, 2016 Nataly Brukhim, Tal Schuster

Theory Questions

1. Let r^* be our target radius. Given the training data S, the algorithm will return the value,

$$\hat{r} = \max_{(x,y)\in S}\{r|x^2+y^2=r^2 \wedge c_{r*}(x,y)=1\}$$

Let r_ϵ defined as follows,

$$r_\epsilon = inf\Big\{r|\underset{(x,y)\sim\mathcal{D}}{P}\big(r\leqslant\sqrt{x^2+y^2}\leqslant r^*\big)=\epsilon\Big\}$$

Let R be the ring defined as follows,

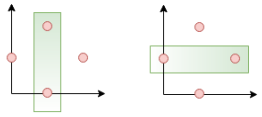
$$R=\{(x,y)|r_\epsilon\leqslant\sqrt{x^2+y^2}\leqslant r^*\}$$

Clearly, from the definition of R, we have $\underset{(x,y)\sim\mathcal{D}}{P}\big((x,y)\in R\big)=\epsilon$. If no point is in the ring R, the error exceeds ϵ. Thus,

$$P(error>\epsilon)=\prod_{i=1}^{m}P\big((x_i,y_i)\notin R\big)=(1-\epsilon)^m<e^{-m\epsilon}<\delta$$

Which implies that the required sample size is $m>\frac{1}{\epsilon}log(\frac{1}{\delta})$.

2. First, we will show that $VCdim(\mathcal{H})\geqslant 4$. Let $S=\{(1,0),(0,1),(2,1),(1,2)\}$ be our sample set which consists of four 2D samples. We need to show that $|\Pi_{\mathcal{H}}(S)|=16$. It can be easily shown that there exists an axis-aligned rectangle for each possible labeling configuration of the samples. Two such possible labeling configurations are demonstrated in the following image:



Thus, the VC-dimension of $\mathcal{H}$ is at least 4.

We will now show that $VCdim(\mathcal{H})<5$. Let $S=\{s_1,s_2,s_3,s_4,s_5\}$ be our sample set which consists of five 2D distinct samples. The labeling configuration in which the most extreme points that lie on the tightest bounding axis-aligned rectangle (one most extreme per edge), are assigned positively. Clearly, there are at most 4 such extreme points on the rectangle. The fifth point is either in the interior of the rectangle or located somewhere on an edge, along which another positively labeled point is located. Hence, by labeling the fifth point negatively we have found a labeling which is inconsistent with any axis-aligned rectangle. Thus, the VC-dimension of $\mathcal{H}$ is at most 4.

3. First, we will show that $VCdim(\mathcal{H})\geqslant n$. Let $S=\{\mathbf{e}_1,...,\mathbf{e}_1\}$ be the set of unit vectors. We need to show that $|\Pi_{\mathcal{H}}(S)|=2^n$, which is easily achieved by setting the indices of T_1 to be those of the positively labeled samples in some labeling, while T_2 can be empty.

We will now show that $VCdim(\mathcal{H})<n+1$. Given $n+1$ vectors, there exists a vector $\mathbf{v}$ that is a linear combination of some other vectors, denoted by $\mathbf{v}_1,...,\mathbf{v}_\ell$. Since any hypothesis h_{T_1,T_2} is a function over only n indices, for any such hypothesis h_{T_1,T_2}, its values for $\mathbf{v}$ and for $\mathbf{v}_1,...,\mathbf{v}_\ell$ are dependent. Specifically, there exists some labeling to the vectors $\mathbf{v},\mathbf{v}_1,...,\mathbf{v}_\ell$ which is inconsistent with any hypothesis. Thus, the VC-dimension of $\mathcal{H}$ is at most n.

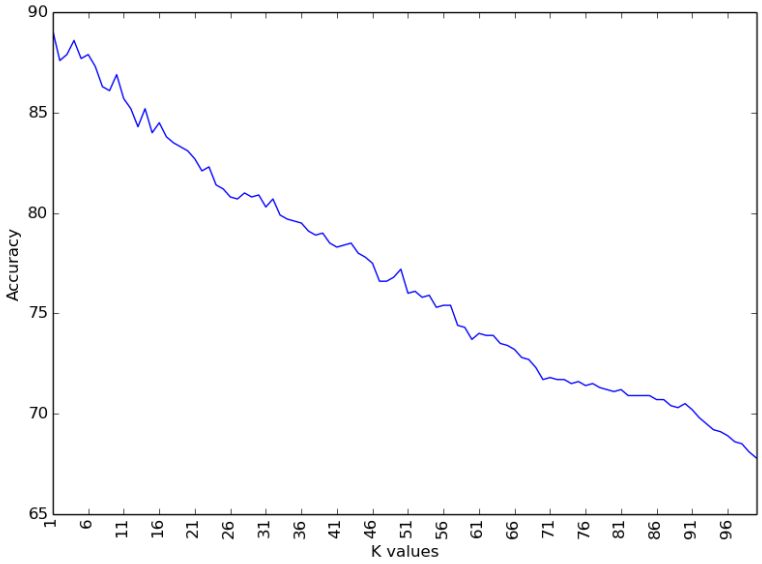
Programming Assignment

1. (a) The code should be run by "python /a/home/cc/students/cs/talschuster/hw/ex1/q1/run.py a"

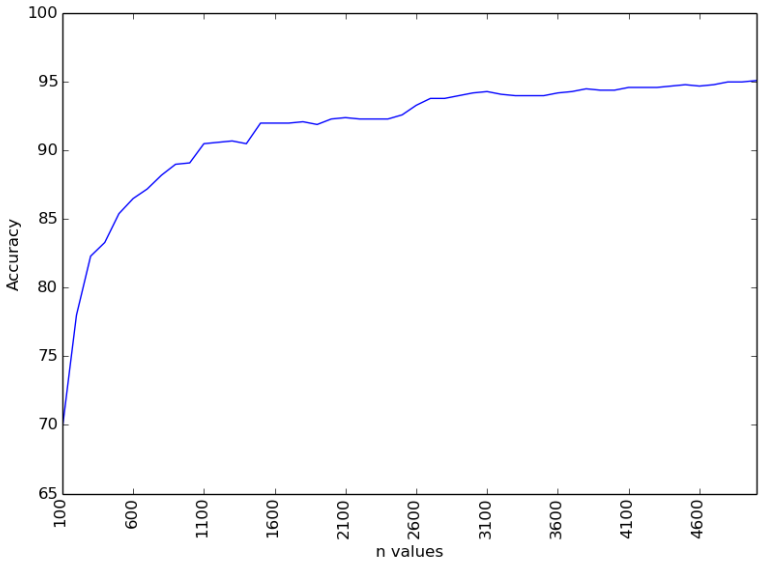
   (b) The code should be run by "python /a/home/cc/students/cs/talschuster/hw/ex1/q1/run.py b"

   The accuracy of the prediction for $k = 10$ is 86.9%. Since there are 10 possible labels for each samples, for a completely random predictor: $P[h(x) = c(x)] = \frac{1}{d} = \frac{1}{10}$. Therefore, we will expect it's accuracy to be 10%.

   (c) The accuracy is gradually decreasing for higher values of k. Best value is $k = 1$. The reason might be that, after choosing the k neighbors, only the amount of labels is determining the prediction with no value to the distance of each sample. Therefore, the single closest sample might be better than several samples samples that are in the general same area.

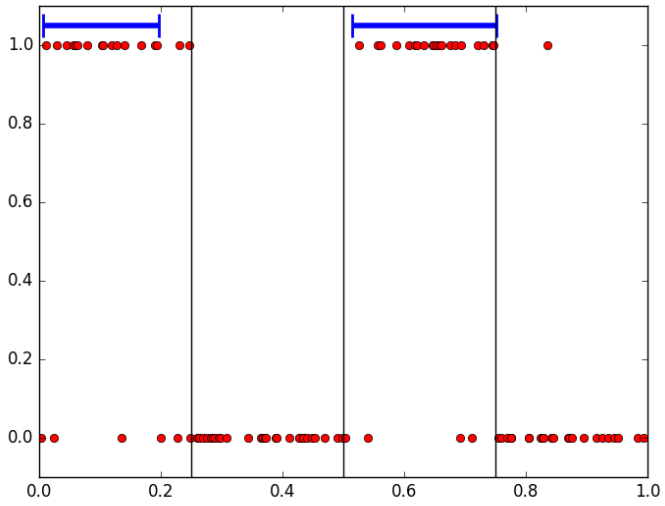


   (d) Using a larger amount of training images improve the accuracy. An intuitive explanation is that by applying k-NN algorithm we relay on the assumption that closer samples have similar labels. By increasing the amount of training samples, more samples are being added to the examined space. Therefore, by using more training samples, the distance between each test sample to its nearest neighbor is equal or lower. Specifically, in the case of 1-NN, the training sample that determines the prediction will be closer to the examined sample and therefore, the prediction will be more accurate.



2. The code for the following questions can be found in:
"/a/home/cc/students/cs/talschuster/hw/ex1/q2/**X**.py", for **X**=a/c/d/e/f.

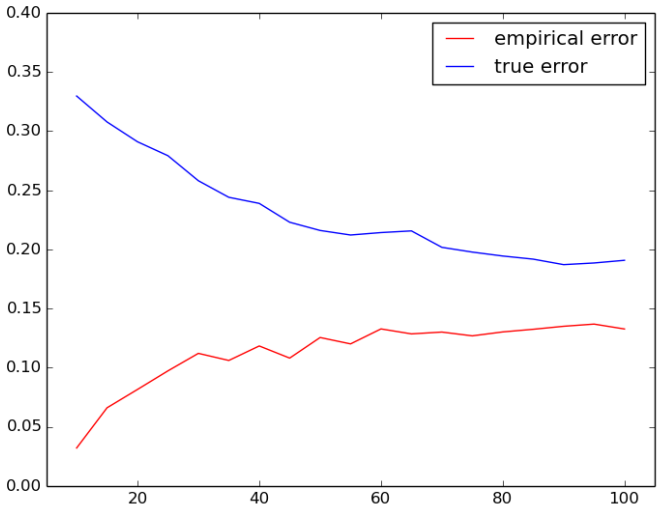
(a) The following image is a plot of 100 samples drawn from D, where the vertical lines are located at $x = 0.25, 0.5, 0.75$, and the best 2 intervals (as found by `find_best_interval`) are shown in blue.



(b) The hypothesis with the smallest error is given by, $\underset{h \in H}{argmin}\{error(h)\} = \underset{h \in H}{argmin}\{\underset{(x,y)\sim D}{P}\big(h(x) \neq y\big)\}$. This hypothesis consists of the intervals $[0, 0.25], [0.5, 0.75]$, and gives the error

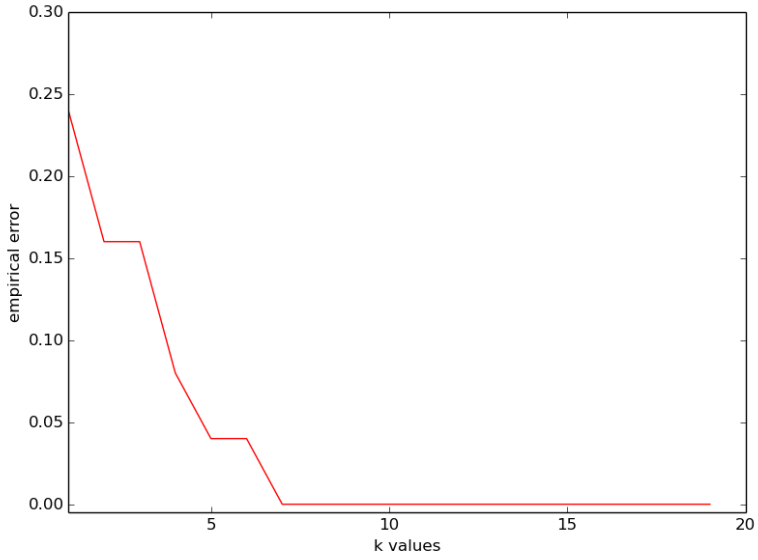
$$\underset{h \in H}{min}\{error(h)\} = \frac{1}{2} \cdot \frac{2}{10} + \frac{1}{2} \cdot \frac{1}{10} = 0.15$$

(c) The following image is a plot of the average empirical and true errors, averaged across the 100 runs, as a function of m. The empirical and true errors are shown in red and blue, respectively.



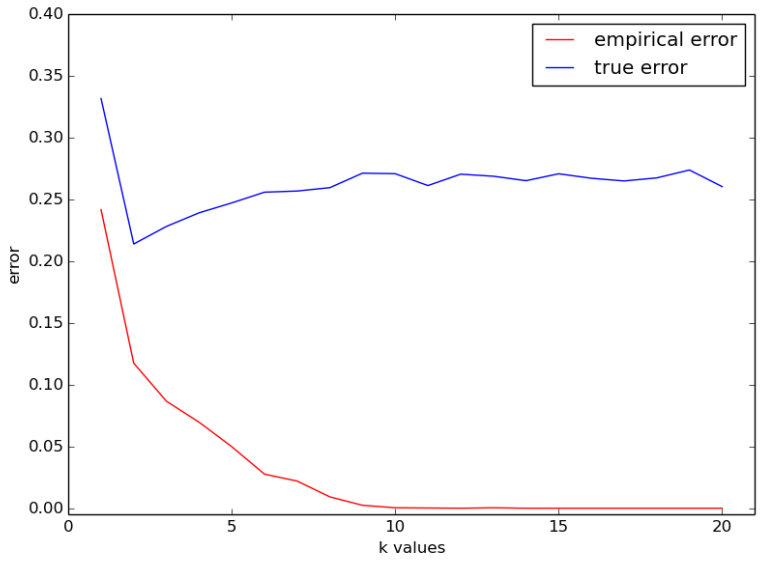
The true error is decreasing as m increases, since by using more data, the ERM algorithm is able to get a better hypothesis which decreases its error over the true distribution. The empirical error increases as m increases since the error is more easily minimized with fewer samples, and vice verca. The empirical error approaches the true error as more data is used.

(d) The following image is a plot of the empirical error as a function of k.



In this setting, $k^* = 7, ..., 20$. This experiment doesn't necessarily mean that the hypothesis with k^* intervals is a good choice, for a few reasons. First, this depends on the size of the data we have used, which might be insufficient to generalize well. Additionally, this is a random draw of samples from the true distribution, and in order to get a more robust result we should have used multiple draws and average the error across runs. Finally, we know that by using too many intervals, applying the ERM algorithm will cause overfitting (e.g. by setting an interval per positive sample), hence choosing the value that achieves the minimal error over the training data is not necessarily a good choice.

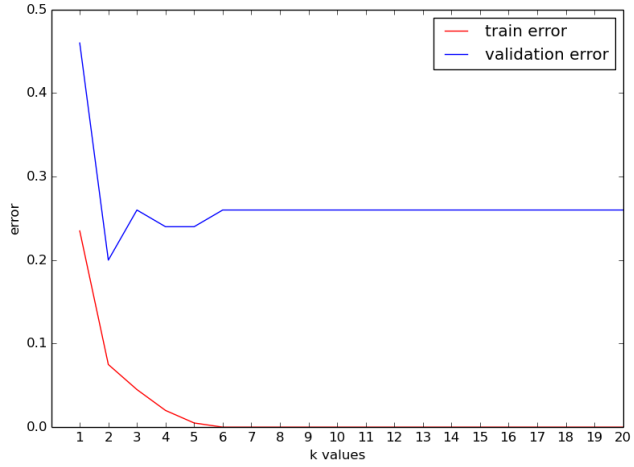
(e) The following image is a plot of the empirical error and true error averaged across runs, each run generated by the above procedure.



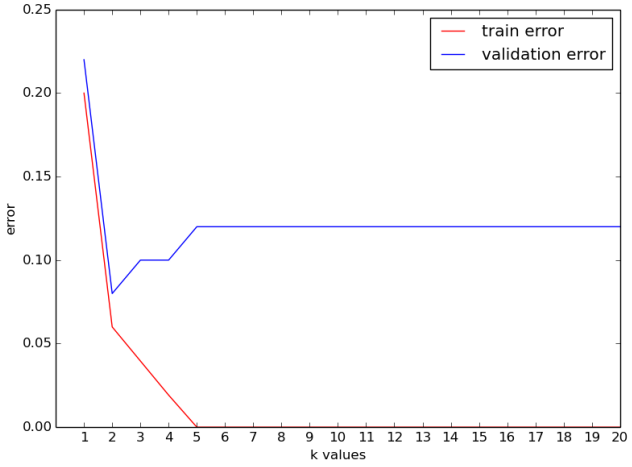
As stated in the previous question, choosing the value that achieves the minimal error over the training data is not necessarily a good choice. This is demonstrated in the above plot in which the minimal true and empirical errors do not agree.

(f) To perform a cross-validation (CV) with the m samples we need to use n samples for validation and train on the other $m-n$ samples. We will repeat this experiment $\frac{n}{m}$ times, each time using different n samples as the validation set. This experiment is repeated for different k values. Since we assume here that we don't know the true distribution, this gives us a good measurement because the validation error is a prediction of how our trained model will perform for new samples. The use of cross-validation allows us to use all the data when choosing the best k value.
Note: the partition to sets needs to be from random indexes since the data here is sorted.

Below are the results for a 5-fold CV (a) and a leave-one-out CV (b). From both we can conclude that the best k to use is 2, which is the same k value that gave the lowest true error on the previous section. The validation error for the leave-one-out is lower since we use more data for training the model.



(a) 5-fold cross-validation



(b) leave-one-out cross-validation