AI 539 HW3

Matthew Pacey

February 19, 2024

1 Experimental Setup

Our problem setting is that we have training and test data that contains two dimensions of features contained in a two-dimensional point, where each feature is normalized to a value between 0 and 1. Each data point is labeled with either a positive or negative class. Our goal is to demonstrate empirical results of a histogram plug-in classifier. The classifier works by dividing the input space into an m by m grid of cells. Of the training points, n are sampled and placed into the cells by their position. The labels in each cell are collected and the majority class dictates the classifier's prediction (a coin flip is used to determine the prediction if the positive and negative label counts match). We are evaluating over the combinations of $m \in \{2, 4, 8, 16\}$ and $n \in \{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. The empirical risk is then calculated by running all test points through each of the classifiers and recording the number of incorrect predictions by the classifier divided by the number of test points.

2 Numerical Experiment Results

A Averages

In Figure 1, we see the average performance of each classifier. Each curve represents dividing the input space into an m by m grid. The values on the x-axis represent increasing n values (in log scale) which represents the number of training points that are selected to build the classifier. The y-axis represents the average empirical risk over 100 trials. The risk is calculated by feeding all of the the test points through the classifier (trained on n training examples) and taking the sum of misclassified points over the total number of test points. The minimum risk (a given constant for this problem) is subtracted from each average to produce the plot.

A key takeaway from Figure 1 is the relative invariance of the m=2 (blue) curve. All m values start with relatively similar average risk values but most other m-value curves led to lower risk as the number of training examples (n) increased. To explain this, we can consider the coarseness of the grid created by m of 2 (4 buckets to fit test points into) versus m of 16 which has 4096 buckets for points. Given how few buckets there are in the case of m=2, there is no virtually no benefit in using more training examples, it is clear that the buckets must be smaller for the classifier to be more effective. We can see the effectiveness of more buckets in the m of 16 (red) line. Also along this line is the diminishing returns of more training examples. In the intermediate m values (m of 4 and 8) both lines are nearly unchanged when increasing the number of training examples twofold (from 10^4 to 10^6). This demonstrates that simply adding more training examples will not lead to perfect classification, there is a limit for every configuration.

B Scatterplot

Figure 2 shows the same trials as Figure 1 but with more detail. Instead of averaging the risk across all trials, each trial is shown on the plot. Each trial has a semi transparent dot (color grouped by m value) so that places where multiple results generated the same risk are colored darker. In addition to plotting all values, the fifth highest and fifth lowest risk values of each m value are connected for each successive n value. With 100 trials, these bounds represent where 90% of the results for that m value fell.

One takeaway from this plot is the relative widths of the bounding lines. As expected, the bounds were wider for lower m values and tighter for higher m values. We can again consider how the coarseness of classifier is dictated by m. With lower m values, there will be more variance in the generated classifier (particularly for lower n values) leading to a larger variance in risk calculated on the test data. Similarly, each curve shows a point of n training examples where the fifth best and worst lines are directly on top of each other. This occurs at 10^3 for m=2, 10^4 for m=4, 10^5 for m=8, and 10^6 for m=16. This reinforces the point that as m continues increasing, there will be less and less variance in the classifier, demonstrating diminishing returns by continuing to add more training data.

C Selected Probability Heatmaps and Classifiers

The plots in Figure 3 demonstrate a portion of the input data and classifiers that were used to achieve the empirical results. The first row is generated using m of 2 and the second row is generated with m of 16. The figures on the left are probability heatmaps of the input data. Darker colors represent areas where one class is dominant (e.g. blue in upper right of each plot) and lighter colored areas represent areas where the classes are more balanced. The two plots on the right side represent how the probabilities on the left plots are translated into classifiers. The majority class is used to pick the prediction. The increase in the m value led to a finer classifier, allowing the m of 16 value to better reflect the curve between the positive and negative classes, whereas the m of 2 will have worse predictions in the border areas because its resolution is too small.

3 Plots

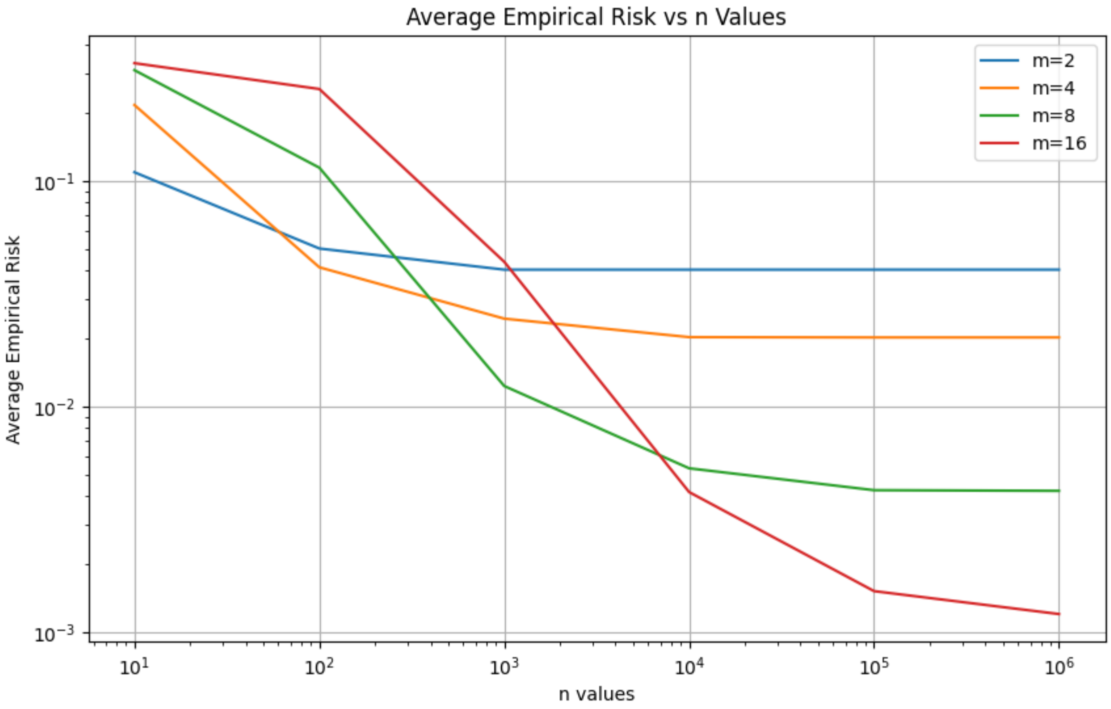

Figure 1: Average Empirical Risks

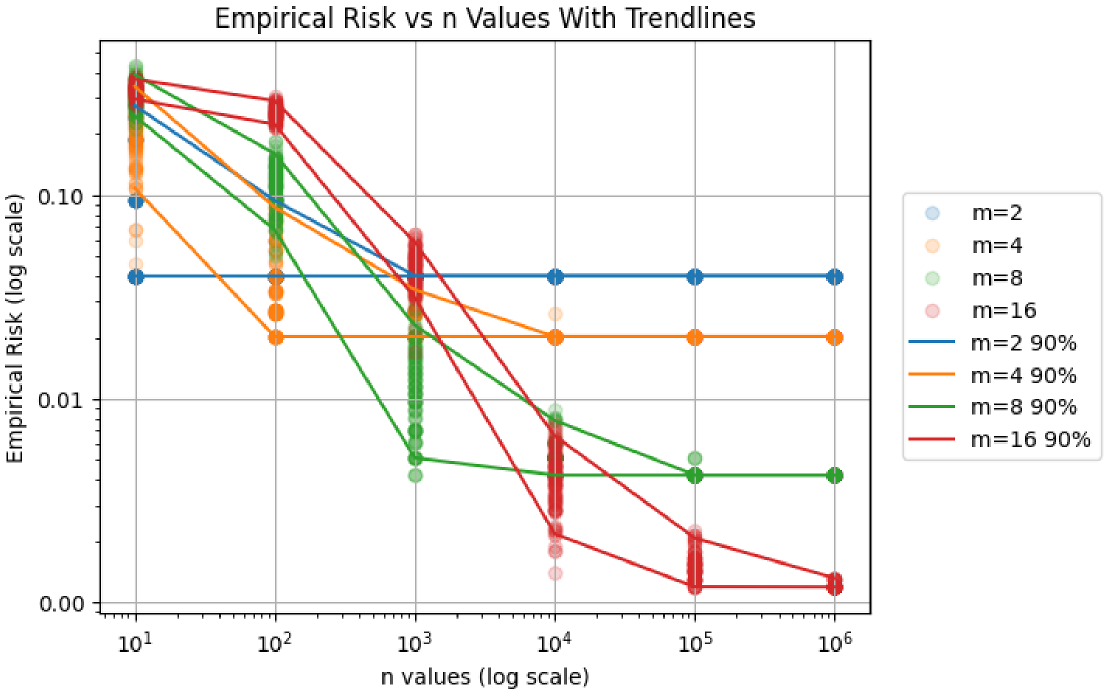

Figure 2: All Empirical Risks With Trendlines

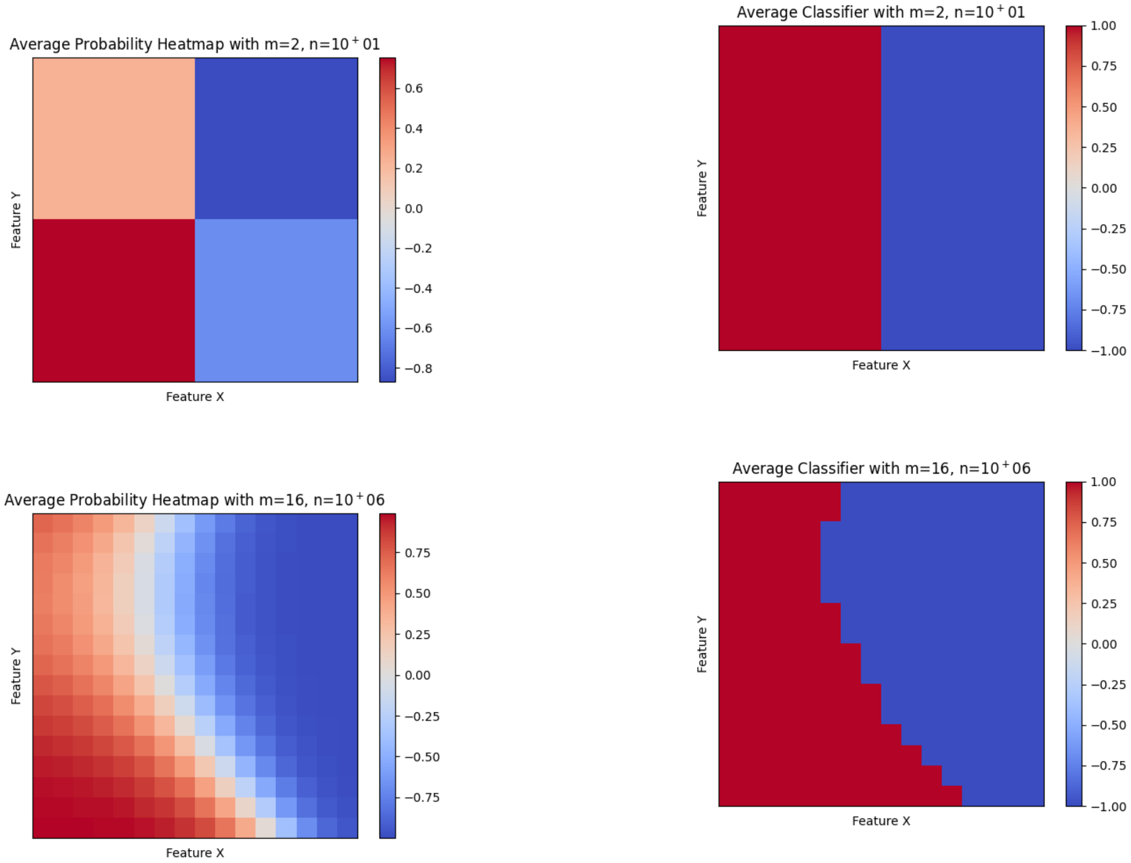

Figure 3: Selected Probability Heatmaps (left) and Classifiers (right)