Homework 1: ML

Selected Solutions

Q3 Considering now $\mathcal{H}_d$, with $d > 2$. Justify an inequality between $R(f^*_{\mathcal{H}_d})$ and $R(f^*_{\mathcal{H}_2})$. Which function $f^*_{\mathcal{H}_d}$ is a risk minimizer in $\mathcal{H}_d$? What is the approximation error achieved by $f^*_{\mathcal{H}_d}$?

In a general case, the inequality is: $R(f^*_{\mathcal{H}_d}) \leq R(f^*_{\mathcal{H}_2})$

For any subspace $\mathcal{H}_d$, with $d > 2$, all functions from $\mathcal{H}_2$ are included in $\mathcal{H}_d$ because you can set the coefficients $b_k, k \in \{3...d\}$ to zero and set b_0, b_1, b_2 accordingly to match the required function from $\mathcal{H}_2$. So the risk minimizer function from the hypothesis space $\mathcal{H}_2$, $f^*_{\mathcal{H}_2}$, is included in $\mathcal{H}_d$, $d > 2$. So the minimzer from $\mathcal{H}_d$, $d > 2$ i.e $f^*_{\mathcal{H}_d}$ has to either further minimze the risk or at worst it matches that from $f^*_{\mathcal{H}_2}$.

The minimizer $f^*_{\mathcal{H}_d}$ is the same as $g(x)$ but it is defined as $b_0 = a_0, b_1 = a_1, b_2 = a_2, b_{3...d} = 0$. And as a result the approximation error is zero again because this is the Bayes predictor.

Q10. Now you can adjust d. What is the minimum value for which we get a "perfect fit"? How does this result relates with your conclusions on the approximation error above?

For d=2 and above there is a perfect fit. This agrees with our conclusions from Q2 and Q3 where the approximation error for $\mathcal{H}_2$ is zero and approximation error for $\mathcal{H}_d(d > 2) \leq \mathcal{H}_2$ (and hence is zero for all of those as well). Basically the true distribution function g belongs to all hypothesis spaces from d=2 onwards. A plot to confirm this programmatically would look something like Figure 1.

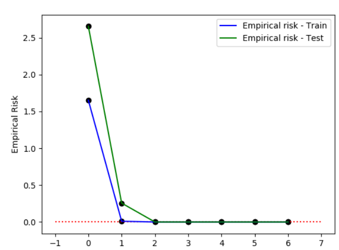
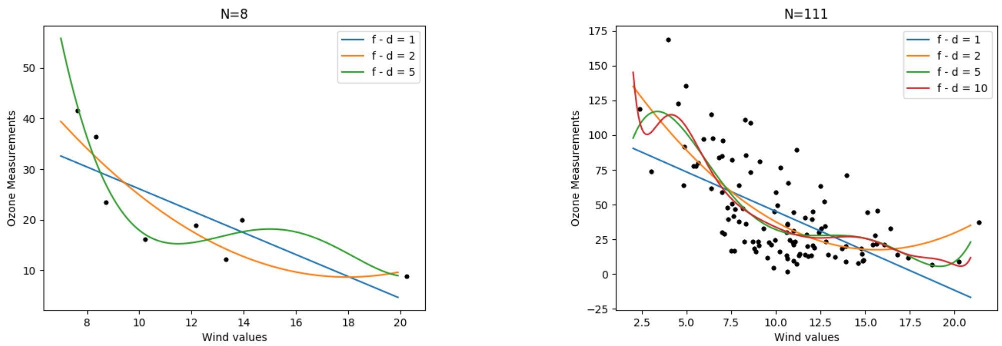


Figure 1: Test for minimum d-value needed for perfect fit. We can see that there is a perfect fit i.e. zero empirical risk, for $d \geq 2$

Q14. Besides from the approximation and estimation there is a last source of error we have not discussed here. Can you comment on the optimization error of the algorithm we



(a) (b)

Figure 2: Scatter plot of data points (each subplot varies N) and estimating functions f to predict ozone value as a function of wind value while varying d

are implementing?

We would expect zero optimization error because our analytical approach finds the exact solution for the least squares optimization. Given the data observed X, we cannot optimize any further than this closed form solution.

Q15. Reporting plots, discuss the again in this context the results when varying N (subsampling the training data) and d.

Here the answer is open ended. As long as you provide plots to view the trends in N and d you should get the points. The general trend (Figure 2) is a decrease in ozone value as wind value increases where a balanced fit is seen with $d = 2$. Within each subsample, as we increase d, we observe more overfitting particularly as d approaches N. From Figure 3, for the same N, a higher d-value usually corresponds to a lower risk value since the more expressive function is better able to represent the train sets. (this is particularly pronounced for smaller N values)

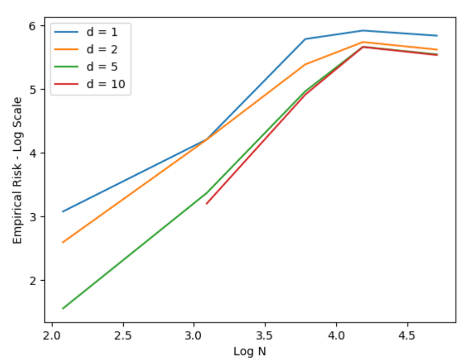


Figure 3: Empirical risk e_t (in the log scale) as a function of log N for d=1, 2, 5, 10