Patterns, Predictions, and Actions

This project introduced me to the ideas and methods of prediction and optimization through a guided reading of the textbook *Patterns, Predictions, and Actions*, written by Moritz Hardt and Benjamin Recht. We began with the most fundamental formulation of prediction as a risk minimization problem. I learned how prediction problems are structured through covariates, labels, prediction functions, loss functions, and risk as expected loss. This framing helped me understand that nearly all prediction tasks, whether regression or classification, can be formulated as minimizing expected loss.

Next, we moved into risk minimization and empirical risk minimization (ERM). I learned that because we are unlikely to be able to observe the true population distribution, we cannot directly compute the true risk. Instead, we minimize empirical risk using observed data. This introduced the important distinction between the ideal population risk minimizer and the practical empirical minimizer. I also learned how the sample size, choice of optimizer, and choice of function class limit the quality of the predictor. This helped me understand the bias-variance tradeoff in a much more intuitive way.

To make these ideas concrete, I implemented several methods using the Breast Cancer Wisconsin dataset from scikit-learn, which contains 569 samples with 30 features and binary labels. First, I built a classifier by assuming that each feature follows a normal distribution within each class. Since this is a strong modeling assumption, I also implemented the ordinary least squares analytical solution, treating the binary labels as 0 and 1. Even though linear regression is not designed specifically for classification, this experiment demonstrated how optimization principles remain consistent across different modeling choices.

Because analytical solutions are not always easy to compute, my mentor introduced me to gradient descent. I implemented full-batch gradient descent to minimize the mean squared error loss and studied how the learning rate affects convergence. By visualizing gradient norms and loss values across epochs, I gained an intuitive understanding of why overly large learning rates lead to divergence while small learning rates result in slow convergence.

We then extended this to stochastic gradient descent and its variations, including shuffling, mini-batch updates, and step-decay learning rates. Through experiments, I observed how stochastic gradient descent introduces randomness into optimization, how mini-batch and learning rate decay help reduce noise in the updates. This hands-on comparison between analytical solutions, gradient descent, and stochastic gradient descent helped me understand modern machine learning better.

Finally, we discussed generalization, which tied everything together. I learned that minimizing training loss alone is not sufficient. The real goal is to perform well on unseen data. This reinforced the importance of model complexity, dataset size, and optimization choices in determining real-world performance.

This DRP project brought together probability, optimization, modeling assumptions, and computation in a way that greatly deepened my understanding of prediction and machine learning.