ICS 635 HW4

April 27, 2019

Problem 1: Equivalence of Common Objectives for Classification

In binary classification we typically model the conditional probability of the N training targets $\{y_n\}_N$ given the inputs $\{x_n\}_N$ using a Bernoulli distribution for each example:

$$p_\theta(y_n|x_n) = \text{Ber}(y_n; \hat{y}_n) = (\hat{y}_n)^y(1-\hat{y}_n)^{1-y}$$

where $y_n \in \{0, 1\}$ is the target for the n-th example and $\hat{y}_n \in [0, 1]$ is a real-valued scalar, interpreted as a probability, which is computed as a function of the n-th input x_n and the model parameters θ.

1. Prove that maximimizing the conditional likelihood of the data is equivalent to minimizing the mean binary cross-entropy loss.

$$\arg\max_\theta p_\theta(\{y_n\}_N|\{x_n\}_N) = \arg\min_\theta \frac{1}{N}\sum_n \left(y_n \log\left(\hat{y}_n\right) + (1-y_n)\log\left(1-\hat{y}_n\right)\right)$$

Problem 2: Describing Machine Learning Methods

Identify at least ten problems or ambiguities with the following (fictional) description that would keep you from reproducing the results. You can write your answer a list of questions.

The Santander Kaggle competition had 200,000 labeled examples for training and 200,000 unlabeled test examples. For model building, the labeled data was split into a training set (80%), validation set (10%), and test set (10%). A neural network was trained using sklearn with the default parameters and early-stopping on the validation set, achieving 90% *accuracy. The hyperparameters were then optimized using Bayesian optimization, and the best hyperparameters achieved a score of 0.99, putting me in first place in the competition. However, the performance on the final test set was slightly worse, suggesting overfitting.*

Problem 3: PCA

Suppose you are given a data matrix X, and you want to perform Principal Component Analysis (PCA) to visualize it in two dimensions.

1. Does centering the data change the principal components of the data? (Warning: This is a tricky question.

2. Does scaling the data change the principal components?

3. Suppose your data consists of samples from two 3-dimensional Gaussian distributions with different means but the same covariance matrices. Describe a scenario in which the two classes would be indistinguishable in the 2D principal-component space.

4. Suppose you train an linear neural network autoencoder on X, with a bottleneck layer of two neurons and mean squared loss. Is this objective function convex? If you train then network, then embed each data point into the 2D latent space and scatter plot it. Will this be equivalent to plotting the data in principal component space?

Problem 4: Gaussian Processes

Suppose you train a Gaussian Process to the model daily total rainfall in Mānoa. To keep things simple, you use a fixed RBF kernel and no noise model.

1. Could you use this to predict *future* rainfall? That is, do you expect it to be better than flipping a coin in predicting whether it will rain?

2. An oracle informs you that there will be no rain two days from now. If you trust the oracle, will this affect your model predictions for whether it will rain tomorrow?

3. The GP model gives predictions for tomorrow's rainfall in terms of a *distribution.* Say you just want a point estimate to display on your surf forecast website. In this case, should you use the Mean Posterior (MP) point estimate of the distribution or the Maximum A Posterior (MAP) estimate?