CS589: Machine Learning - Fall 2025

Homework 1: Classification

Assigned: Thursday, September 10. Due: Wednesday, September 17 at 8:00pm

1. (*10 points*) **KNN and Missing Data:** Missing data is a common problem in real-world applications of machine learning methods. In the classification case, different feature vectors $\mathbf{x}_i = [\mathbf{x}_{i1}, ..., \mathbf{x}_{iD}] \in \mathbb{R}^D$ can have missing values on any subset of the dimensions. We can keep track of which dimensions have valid data and which have missing data using a binary mask vector $\mathbf{m}_i = [\mathbf{m}_{i1}, ..., \mathbf{m}_{iD}] \in \{0, 1\}^D$. For example, if $\mathbf{x}_1 = [5.0, ?, 1.0]$, then $\mathbf{m}_1 = [1, 0, 1]$. Explain how you could modify the standard KNN classifier so that it can be applied to a data set with missing values represented as a set of tuples $(y_i, \mathbf{x}_i, \mathbf{m}_i)$.

Answer: To modify the standard KNN classifier for missing data, we can adjust the distance calculation. Each feature vector $\mathbf{x}_i$ has an associated binary mask $\mathbf{m}_i$, where $m_{i,d} = 1$ if the d-th feature is present and 0 if it is missing. When computing the distance between a new point $\mathbf{x}$ with mask $\mathbf{m}$ and a training point $\mathbf{x}_i$ with mask $\mathbf{m}_i$, we only compare features that are present in both vectors.

The modified Euclidean distance is:

$$d'(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{d=1}^{D} m_d \cdot m_{i,d} \cdot (x_d - x_{i,d})^2}.$$

Here, the product $m_d \cdot m_{i,d}$ equals 1 only if both $\mathbf{x}$ and $\mathbf{x}_i$ have valid values for dimension d, otherwise the term is ignored.

After computing this modified distance for all training points, the rest of the KNN algorithm remains unchanged: we select the K nearest neighbors and predict the class by majority vote.

One limitation of this method is that if many features are missing, the comparison may be based on very few dimensions, which can reduce accuracy. An alternative approach is to impute missing values, for example by replacing them with the mean of the feature, but this may introduce bias into the data.