Introduction to Machine Learning - HW1

Professors: Abolghasemi & Erabi
Student: Mohamad Mahdi Samadi
Student ID: 810101465

Q1.

a. Why do we use cross-validation? Explain at least two of its variations.

a.1. Reasons to use CV:

1. Helps to stop overfitting and such problems in the model by giving a more accurate picture of how the model performs.
2. CV uses all the dataset as both train and test data at different times.
3. CV provides us with more than one result metric for each model, so we would be more confident about the performance of various types of models and choose the better model. By examining the results we'll know which model is more consistent over different parts of the data. The model with more average accuracy and less std on the accuracy is probably the better one.

a.2. CV Variations:

1. K-fold CV: The whole dataset is partitioned in k parts of equal size and each partition is called a fold. K can be any positive integer number. We will train a model k times, at time t (1 <= t <= k) the t'th fold is used for validation and other K-1 folds are used for training the model.
2. Stratified CV: as you might have noticed, k-fold CV can't be used for imbalanced datasets because data is split into k-folds with a uniform probability distribution. Assume 80% of a dataset includes information of male samples and the rest is of females. Using k-fold misleads us to wrong conclusions. Stratified CV is an improved version of the k-fold CV technique. Although it too splits the dataset into k equal folds, each fold has the same ratio of instances of target variables (gender in the mentioned example).

b. Show that multiplying the value of each dimension in a coefficient (where it is a non-zero real number) will lead to a distance metric with the same properties as the Euclidean distance.

$$d(x, y) = \sqrt{\sum_{k=1}^{d} (x_k - y_k)^2}$$

$$d'(x,\ y)\ =\ \sqrt{\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2}\ =\ \sqrt{\sum_{k=1}^{d}a_k{}^2(x_k\ -\ y_k)^2}$$

Let’s check the properties one by one.

1. $d'(x,\ y)\ >\ 0$:

   $(a_k x_k\ -\ a_k y_k)^2\ >\ 0\ \rightarrow$

   $$\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2\ >\ 0\ \rightarrow \sqrt{\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2}\ >\ 0$$

2. $d'(x,\ x)\ =\ 0$

   $(a_k x_k\ -\ a_k x_k)^2\ =\ 0\ \rightarrow$

   $$\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2\ =\ 0\ \rightarrow \sqrt{\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2}\ =\ 0$$

3. $d'(x,\ y)\ =\ d'(y,\ x)$

   $$(a_k x_k\ -\ a_k y_k)^2 = (a_k y_k\ -\ a_k x_k)^2\ \rightarrow \sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2\ =\ \sum_{k=1}^{d}(a_k y_k\ -\ a_k x_k)^2$$

   $$\sqrt{\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2}\ =\ \sqrt{\sum_{k=1}^{d}(a_k y_k\ -\ a_k x_k)^2}\ \rightarrow\ d'(x,\ y)\ =\ d'(y,\ x)$$

4. $d'(x,\ z) \le d'(x,\ y)\ +\ d'(y,\ z)$ (the triangle inequality)

$$d'(x,\ y)\ = \sqrt{\sum_{k=1}^{d}(a_k x_k\ -\ a_k y_k)^2}\ =\ \sqrt{\sum_{k=1}^{d}a_k{}^2(x_k\ -\ y_k)^2}$$

$$\sqrt{\sum_{k=1}^{d}a_k{}^2(x_k\ -\ z_k)^2}\ \le \sqrt{\sum_{k=1}^{d}a_k{}^2(x_k\ -\ y_k)^2}\ +\ \sqrt{\sum_{k=1}^{d}a_k{}^2(y_k\ -\ z_k)^2}\quad (I)$$

Let’s square both sides to get rid of the radicals. First let’s do it for the right side

$$A\ =\ \sum_{k=1}^{d}a_k{}^2(x_k\ -\ y_k)^2 + \sum_{k=1}^{d}a_k{}^2(y_k\ -\ z_k)^2 +\ 2\sqrt{\sum_{k=1}^{d}a_k{}^2(x_k\ -\ y_k)^2 \times \sum_{k=1}^{d}a_k{}^2(y_k\ -\ z_k)^2}$$

Cauchy-Schwarz Inequality: $|\langle u, v\rangle| \leq \|u\| \cdot \|v\|$. Knowing that, we'll rewrite the term inside the radical.

$$\sum_{k=1}^{d} a_k^{2}(x_k - y_k)^2 \times \sum_{k=1}^{d} a_k^{2}(y_k - z_k)^2 \geq [\sum_{k=1}^{d} \alpha_k^{2}(x_k - y_k)(y_k - z_k)]^2$$

$$2\sqrt{\sum_{k=1}^{d} a_k^{2}(x_k - y_k)^2 \times \sum_{k=1}^{d} a_k^{2}(y_k - z_k)^2} \geq 2|\sum_{k=1}^{d} \alpha_k^{2}(x_k - y_k)(y_k - z_k)|$$

$$2|\sum_{k=1}^{d} \alpha_k^{2}(x_k - y_k)(y_k - z_k)| \geq 2\sum_{k=1}^{d} \alpha_k^{2}(x_k - y_k)(y_k - z_k) \quad (II)$$

$$A \geq \sum_{k=1}^{d} a_k^{2}(x_k - y_k)^2 + \sum_{k=1}^{d} a_k^{2}(y_k - z_k)^2 + 2\sum_{k=1}^{d} \alpha_k^{2}(x_k - y_k)(y_k - z_k)$$

$$A \geq \sum_{k=1}^{d} a_k^{2}[(x_k - y_k)^2 + (y_k - z_k)^2 + 2(x_k - y_k)(y_k - z_k)]$$

$$A \geq \sum_{k=1}^{d} a_k^{2}[x_k^{2} + z_k^{2} - 2x_k z_k] = \sum_{k=1}^{d} a_k^{2}(x_k - z_k)^2 = d'(x, z)^2$$

$$A = (d'(x, y) + d'(y, z))^2 \geq d'(x, z)^2 \rightarrow |d'(x, y) + d'(y, z)| \geq |d'(x, z)|$$

Earlier we showed that $d'(a, b)$ is non-negative. So:

$$d'(x, y) + d'(y, z) \geq d'(x, z)$$