Distance metrics in decision trees

Distance metrics for samples:

Cosine similarity: *Cosine similarity is defined by the following equation:*

$$s(x, y) = \frac{xy^T}{||x|| \cdot ||y||}$$

It is called cosine similarity because it potrays the cosine value of the angle between two vectors, which is the dot product over the product of the modules. The value of the similarity ranges from -1 to 1, 1 meaning the two vectors are the same, -1 meaning the two vectors are opposite, and 0 meaning the two vectors are independent of each other. There are a few extensions on the idea of cosine similarity. One of the extensions is Tonimoto index, which is defined by the following equation:

$$T(x, y) = \frac{xy^T}{||x||^2 + ||y||^2 - xy^T}$$

Tonimoto index not only potrays the angle between the two vectors, but takes the length of the two angles into consideration as well. However, the cosine similarity is variant to shifts. And if we want to take shifts out from our description of similarity, we can use Pearson correlation which is defined as:

$$s(x, y) = \frac{(x - \bar{x})(y - \bar{y})^T}{||x - \bar{x}|| \cdot ||y - \bar{y}||}$$

Pearson correlation is can also take the form of the following equation:

$$\rho_{x,y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Minkowski distance: The Minkowski distance of two vectors is defined as:

$$s(x, y) = (\sum_i |x_i - y_i|^p)^{1/p}$$

There are three special conditions for Minkowski distance. When p equals 1, the Minkowski distance is Manhattan distance; when p equals 2, the Minkowski distance is Euclidean distance; when p goes to infinite, the Minkowski distance is Chebyshev distance. The shortcoming of Minkowski distance is that it ignores the scale of each feature dimension and the correlation between different features.

Mahalanobis Distance: If we want to get rid of the effect of feature scale and correlations, we can use Mahalanobis distance which is defined as:

$$s(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

In which S is the covariance matrix. We can see that if the covariance matrix is a unit matrix, which means the features are independent of each other, the Mahalanobis distance will be the same as Euclidean distance.

Distance metrics for binery trees and decision trees:

The currently adopted distance metrics in USPORF is the proximity matrix, which is defined as:

$$D = \frac{L_{i,j}}{S_{i,j}}$$

In which, $L_{i,j}$ *is the count of sample i and sample j fall into the same leaf node, and* $S_{i,j}$ *is the count of all the trees that takes sample i and sample j in the bootstrap sampling.*

There are two other metrics that may be able to potray the distance between two nodes, the first one is the number of edges that may take to go from one node two another, and the other one is the number of back tracing it will take for two samples to land in the same branch.