Lecture 2(Trees)

The lecture introduces and recaps Decision Trees, Random Forests, CART algorithm for supervised tasks, eventually introducing density forests for unsupervised learning(density estimation)
Density estimation is about estimating the unobserved underlying generative model given some observed unlabeled data points. More formally, we want to learn the density $p(v)$ which has generated the data. In a way, the problem of density estimation is closely related to that of data clustering.

Problem: given a set of unlabeled observations we wish to estimate the latent probability density function from which such data has been generated.

What is a density forest? - Density forests can be understood as random forests for density estimations. Instead of splitting examples at each leaf, each leaf models a Gaussian distribution. In this sense, a density forest can be seen as a special case of Gaussian mixture models using hard assignments instead of soft assignments.

Density forest is a collection of randomly trained clustering trees. The tree leaves contain simple prediction models such as Gaussians. Density forest is a generalization of GMMs with two differences:

1. Multiple hard clustered data partitions are created, one by each tree.
2. The forest-based probability is a combination of tree-based probabilities. Each point is explained by multiple clusters.

Questions

1. How can we use RF for sampling from the density?
   a) Grow forest using training data
   b) Randomly select one tree
   c) Randomly move from root to left with probability of moving left being

$$P(left) = \frac{s_L}{s_L + s_R}$$

   d) Randomly sample from covariance in the leaf
   f) Repeat step 4 until the drawn sample is inside the leaf region