1 Trees, tree metrics

Exercise 1 *Show that any tree T is either a star tree or it has two interior vertices v_1 and v_2, each of which is adjacent to exactly one non-leaf vertex.*

Exercise 2 *Using the previous exercise, show that if a tree T has five or more vertices and no vertex of degree 2, then T has at least two disjoint pairs of leaves that form cherries of T.*

Exercise 3 *Let T be a binary undirected tree with n leaves. Show that the number of its edges is $2n-3$.*

Exercise 4 *Suppose that $D = [d_{ij}]$ is a tree metric on $[m] := \{1, \ldots, m\}$ on a binary tree and fix $r \in [m]$.*

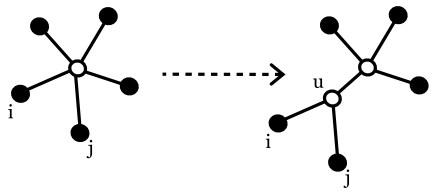
(a) Show that any pair $i, j \in [m] \setminus \{r\}$ that minimizes

$$d_r(i,j) \;:=\; \frac{1}{2}(d_{ij} - d_{ir} - d_{jr})$$

is adjacent to a common vertex u.

(b) Provide formulas to calculate the distance from u to any element in $[m]$.

(c) This means that the underlying tree can be easily recovered recursively: Start with a star-tree. Find the first pair i, j as above and update the tree as follows:



Then proceed recursively replacing i, j with u. Argue why the final result does not depend on the choice of r.

Note: This gives a natural interpretation of the Neighbor-Joining method briefly mentioned in Lecture 1. If D is not a tree metric, the procedure may depend on the choice of r and so we average this choice out.

Exercise 5 *Suppose that D is a tree metric on a tree $\mathcal{T}$ with positive edge lengths d_e. Show that the set of splits A/B of the set of leaves such that*

$$d_{ij} + d_{kl} \;<\; d_{ik} + d_{jl} \qquad \text{for all } i,j \in A, k,l \in B$$

is precisely the set of $\mathcal{T}$-splits.

Exercise 6 *Consider the space of phylogenetic oranges on a tripod tree. Show carefully that the model is defined by the following inequalities: $\rho_{12}, \rho_{13}, \rho_{23} \in [0,1]$, and $\rho_{12} \geq \rho_{13}\rho_{23}$, $\rho_{13} \geq \rho_{12}\rho_{23}$, $\rho_{23} \geq \rho_{12}\rho_{13}$.*

2 Models on graphs and on trees

Exercise 7 *Suppose X, Y are two binary random variables with joint distribution $P = [p_{ij}]$ for $i,j = 0,1$. Show that $\operatorname{cov}(X,Y) = \det P$.*

Exercise 8 *Suppose that $X_1, \ldots, X_n$ are binary variables that form a Markov chain. Find the expression for $\operatorname{cov}(X_1, X_n)$ in terms of the transition matrices representing the conditional distributions of X_i given X_{i-1} for $i = 2, \ldots, n$.*

Exercise 9 *Consider a simple example of the binary non-homogeneous Markov chain $X_1, \ldots, X_n$ as in the previous exercise. Suppose that transition probabilities are symmetric in the sense that*

$$P^{(i)} \;=\; \begin{bmatrix} 1-p_i & p_i \\ p_i & 1-p_i \end{bmatrix}.$$

This process can be viewed as one that flips between two states with probability p_i that may vary between time steps. This process has the uniform stationary distribution $\pi = [0.5, 0.5]$. Show that the n-step transition matrix $P^{(1)} \cdots P^{(n)}$ is also symmetric and its off-diagonal entry is given by

$$p = \frac{1}{2}\left(1 - \prod_{i=1}^{n}(1-2p_i)\right).$$

Exercise 10 (Symmetric Ising model) *Let $T = (V,E)$ be an undirected tree. Consider a family of probability distributions on $\mathcal{X} = \{-1,1\}^n$ given in the following form*

$$p(x) \;=\; \frac{1}{Z(J)} \exp\{\sum_{uv \in E} J_{uv} x_u x_v\},$$

where $Z(J)$ is the normalizing constant

$$Z(J) \;=\; \sum_{x\in\mathcal{X}} \exp\{\sum_{uv\in E} J_{uv}x_u x_v\}$$

often called the partition function. Find a closed form expression for the normalizing constant.

Exercise 11 *Suppose that (X, Y, Z) are jointly Gaussian. Show that $X \perp\!\!\!\perp Y|Z$ if and only if the correlations satisfy the equation* $\mathrm{corr}(X, Y) = \mathrm{corr}(X, Z)\mathrm{corr}(Y, Z)$.

Exercise 12 *We will now generalize the above exercise as follows. Suppose that (X, Y, Z) admit a joint distribution and the conditional expectations $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$ are linear functions (of Z). Show that $X \perp\!\!\!\perp Y|Z$ if and only if the correlations satisfy* $\mathrm{corr}(X, Y) = \mathrm{corr}(X, Z)\mathrm{corr}(Y, Z)$.

Exercise 13 (Linear latent models) *Suppose that all X, Y are discrete random variables with $k+1$ states. We encode them as $\mathbb{R}^k$-valued variables with values $\mathbf{0}$, $\mathbf{e}_1$, ..., $\mathbf{e}_k$. Derive the expression for $\mathbb{E}[X|Y]$ and show it is an affine function of Y.*

3 Tree inference

Exercise 14 *Consider the fully observed tree model on T, where all the vertices represent variables with a finite number of states. Suppose that a random sample of size n is observed and let $\hat{p}$ be the sample distribution (sample proportions). The maximum likelihood estimator satisfies*

$$p(y;\hat{\theta}) \quad = \quad \prod_v \hat{p}_v(y_v) \prod_{uv\in E(T)} \frac{\hat{p}_{uv}(y_u, y_v)}{\hat{p}_u(y_u)\hat{p}_v(y_v)}.$$

Show that the log-likelihood at the maximum likelihood estimator can be rewritten as

$$n\sum_v \sum_{y_v} \hat{p}_v(y_v)\log\hat{p}_v(y_v) + \; n\sum_{u-v\in E(T)} \boldsymbol{I}_{\hat{p}}(Y_u, Y_v)$$

proving correctness of the Chow-Liu alorithm.