Lecture 8: January 20

Lecturer: Donlapark Pornnopparath

We continue with the score-based approach.

8.1 Maximum likelihood parameters

We desire to find a graph with the highest maximum likelihood; the score function of how well a graph G fits data D is defined by

$$s_L(G|D) = LL(\hat{\theta}_G|D),$$

where LL is the log-likelihood function and $\hat{\theta}_G$ is the maximum likelihood parameters for G. To see the connection between the likelihood score and the information theory, we consider a simple example where G consists of two nodes X and Y. The graph G_0 in which two nodes are independent has the score

$$s_L(G_0|D) = \sum_i \log \hat{\theta}_{x_i} + \log \hat{\theta}_{y_i}.$$

Another graph $G_1 : X \to Y$ has the log-likelihood score of

$$s_L(G_1|D) = \sum_i \log \hat{\theta}_{x_i} + \log \hat{\theta}_{y_i|x_i}.$$

We would like to know how far away these two scores are from each other

$$\begin{aligned} s_L(G_1|D) - s_L(G_0|D) &= \sum_i \log \hat{\theta}_{y_i|x_i} - \log \hat{\theta}_{y_i} \\ &= \sum_i \log \hat{\theta}_{y_i|x_i} / \hat{\theta}_{y_i} \\ &= \sum_{x,y} n(x,y) \log \hat{\theta}_{y|x} / \hat{\theta}_y \\ &= n \sum_{x,y} \frac{n(x,y)}{n} \log \hat{\theta}_{y|x} / \hat{\theta}_y. \end{aligned}$$

We have already computed that $\hat{\theta}_{y|x} = n(x,y)/n(y) = \widehat{P}(y|x)$ and $\hat{\theta}_y = n(y)/n = \widehat{P}(y)$. Therefore, the difference can be written as

$$s_L(G_1|D) - s_L(G_0|D) = n \sum x, y \widehat{P}(x,y) \log \frac{\widehat{P}(y|x)}{\widehat{P}(y)} = n \cdot I_{\widehat{P}}(X,Y),$$

where $I_{\widehat{P}}(X,Y)$ is the *mutual information* between X and Y, which measures the strong dependency of Y on X. Thus, stronger dependency implies stronger preference of model G_1 over G_0. One can generalize and obtain the following result.

Proposition 8.1. Let G be a graph and $\widehat{P}$ the empirical distribution from data D. Then the likelihood score decomposes as follows:

$$s_L(G|D) = n\sum_{i=1}^{m} I_{\widehat{P}}(X^i, \mathrm{Pa}_{X^i}) - n\sum_{i=1}^{m} H_{\widehat{P}}(X^i),$$

where $H_{\widehat{P}}(X^i)$ is the *entropy* of X^i given by

$$H_{\widehat{P}}(X^i) = \sum_{x^i} \widehat{P}(x^i) \log \frac{1}{\widehat{P}(x^i)}.$$

Here, $H_P(X)$ measures the unpredictability of the outcome; for example the entropy of casting a die ($\log 6$) is higher than that of tossing a coin ($\log 2$). Since the entropy term is the same regardless of the network structure, the log-likelihood score measures the strength of the dependency between each variable and its parents. Despite of this intuitive interpretation, the log-likelihood score always prefers more complex network and hence cause the model to overfit the data. To see this, we go back to the basic example of G_0 versus G_1. Since the mutual information is non-negative it follows that

$$\begin{aligned} s_L(G_1|D) - s_L(G_0|D) &= n \cdot I_{\widehat{P}}(X, Y) \geq 0 \\ s_L(G_1|D) &\geq s_L(G_0|D). \end{aligned}$$

Thus the score always prefer G_1 over G_0 regardless of the training data. This is also the case for more complex graphs; given any three variables X, Y and Z, the mutual information has the following property:

$$I_P(X, Y \cup Z) \geq I_P(X, Y)$$

with equality holding only if $Z \perp\!\!\!\perp X|Y$. Intuitively, this follows from the fact that adding Z almost always give us more information, unless we observe the exact independence from the data. However, this is usually not the case due to noises, and it is most likely that the maximum likelihood graph will be the complete one. This is an example of *overfitting* the training data (or more precisely, its empirical distribution). In view of this, we add a *regularizing term* to the score function

$$s_L(G|D) = LL(\hat{\theta}_G|D) - \phi(n)\|G\|,$$

where n is the number of samples in the data, $\|G\|$ is the number of parameters in G and ϕ is an increasing function. For AIC, $\phi(n) = 1$ and for BIC, $\phi(n) = \log(n)/2$; the former function penalizes more complex graphs, while the reasoning behind the latter will be touched on in the next section.

8.2 Bayesian score

We now look for an alternative that avoids overfitting the training data. In this section, we examine a score which is based on the Bayesian approahc; since we are uncertain about the network and its parameters, we place the following priors:

- a structure prior $P(G)$ which gives the probability to each graph structure
- a parameter prior $P(\theta_G|G)$ which gives the probability to each possible value of parameters given a graph G.

By Bayes' rule, we have

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)},$$

Note that the denominator does not depend on the graph structure. Thus we define the *Bayesian score* as

$$s_B(G|D) = \log P(D|G) + \log P(G).$$

We will focus on the *marginal likelihood* term

$$P(D|G) = \int_{\Theta_G} P(D|\theta_G, G)P(\theta_G|G)\ d\theta_G.$$

Here, G is a fixed network, $P(D|\theta_G|G)$ is the likelihood of data given the parameters θ_G and $P(\theta_G|G)$ is the prior distribution of these parameter.

8.2.1 A simple graph

We demonstrate here how to compute the marginal likelihood for a simple case. Let us go back to the coin tossing problem with results H, T, T, H, H. The maximum likelihood value is

$$P(D|\hat{\theta}) = \left(\frac{n(H)}{n}\right)^{n(H)} \left(\frac{n(T)}{n}\right)^{n(T)} = \left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^2 \approx 0.035.$$

Alternatively, we use the Bayesian approach and assume a Beta prior. To compute the marginal likelihood, we utilize the chain rule

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)\ldots P(x_n|x_1, x_2, \ldots, x_{n-1}).$$

Recall that the probability of a prediction is given by

$$P(x_{l+1}|x_1, x_2, \ldots, x_l) = \frac{n_l(H) + \alpha_1}{l + \alpha},$$

where $n_l(H)$ is the number of heads in the first l trials. Therefore,

$$P(x_1, x_2, \ldots, x_5) = \frac{\alpha_1}{\alpha} \cdot \frac{\alpha_0}{1+\alpha} \cdot \frac{1+\alpha_0}{2+\alpha} \cdot \frac{1+\alpha_1}{3+\alpha} \cdot \frac{2+\alpha_1}{4+\alpha}.$$

If we let $\alpha_0 = \alpha_1 = 1$, then we have

$$P(x_1, x_2, \ldots, x_5) = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{4} \cdot \frac{2}{5} \cdot \frac{3}{6} = \frac{[1 \cdot 2 \cdot 3] \cdot [1 \cdot 2]}{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = \frac{12}{720} \approx 0.017.$$

We can see that the MLE estimator gives a higher probability since it was designed to be an optimal fit to the current outcome, while the one given by the marginal likelihood is smaller because of the prior belief that the number of heads and tails should be equal. With prior $\text{Beta}(\alpha_1, \alpha_0)$ in general, we have

$$\begin{aligned} P(x_1, x_2, \ldots, x_n) &= \frac{\prod_{i=0}^{n(H)-1}(\alpha_1 + i)\prod_{i=0}^{n(T)-1}(\alpha_0 + i)}{\prod_{i=0}^{n-1}(\alpha + i)} \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \cdot \frac{\Gamma(\alpha_1 + n(H))}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_0 + n(T))}{\Gamma(\alpha_0)}. \end{aligned}$$

If instead we have a Dirichlet prior with multinomial distribution over the space $\{1, 2, \ldots, m\}$, then a similar formula holds:

$$P(x_1, x_2, \ldots, x_n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \cdot \prod_{k=1}^{m} \frac{\Gamma(\alpha_k + n(k))}{\Gamma(\alpha_k)}.$$