Lecture 11: Maximum Entropy

- Maximum entropy principle
- Maximum entropy distribution
- Applications

Dr. Yao Xie, ECE587, Information Theory, Duke University

Berger's Burger

Item	Price	Calories
Burger	\$1	1000
Chicken	\$2	600
Fish	\$3	400
Tofu	\$8	200

- A graduate student's daily average meal cost $= \$2.5$
- What is the frequency that each item being ordered?

- $p(B) + p(C) + p(F) + p(T) = 1$
- $\$1p(B) + \$2p(C) + \$3p(F) + \$8p(T) = \$2.5$
- Still cannot determine the frequencies uniquely ...

Maximum entropy principle

- Maximum entropy principle arose in statistical mechanics
- If nothing is known about a distribution except that it belongs to a certain class
- Distribution with the largest entropy should be chosen as the default
- Motivation:
  - Maximizing entropy minimizes the amount of prior information built into the distribution
  - Many physical systems tend to move towards maximal entropy configurations over time

Physics

- Temperature of a gas corresponds to the average kinetic energy of the molecules in the gas

$$\sum_i p_i \frac{1}{2} v_i^2 m_i$$

- Distribution of velocities in the gas at a given temperature

- this distribution is the maximum entropy distribution under the temperature constraint: Maxwell-Boltzmann distribution

- corresponds to the macrostate that has the most micro states

Formulation

- Maximize entropy

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

- Subject to

$$p_i \geq 0 \tag{1}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{2}$$

$$\sum_{i=1}^{n} p_i r_{ij} = \alpha_j, \ \text{for } 1 \leq j \leq m \tag{3}$$

Maximum entropy distribution

- Form Lagrangian

$$J(p) = -\sum_{i=1}^{n} p_i \log p_i + \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) + \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} p_i r_{ij} - \alpha_j \right)$$

- Take derivative with respect to p_i: $-1 - \log p_i + \lambda_0 + \sum_{j=1}^{m} \lambda_j r_{ij}$

- Set this to 0, and solution is *maximum entropy distribution*

$$p_i^* = \frac{e^{\sum_{j=1}^{m} \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

- $\lambda_0, \lambda_1, \ldots, \lambda_m$ are chosen such that $\sum_i p_i^* = 1$, and $\sum_i p_i^* r_{ij} = \alpha_j$.

Burger's problem

- $p^*(B) = e^{\lambda_0 - 1 + \lambda_1}$, $p^*(C) = e^{\lambda_0 - 1 + 2\lambda_1}$, $p^*(F) = e^{\lambda_0 - 1 + 3\lambda_1}$, $p^*(T) = e^{\lambda_0 - 1 + 8\lambda_1}$

- $p(B) + p(C) + p(F) + p(T) = 1$

- $p(B) + 2p(C) + 3p(F) + 8p(T) = 2.5$

- Solution: $\lambda_0 = 1.2371$, $\lambda_1 = 0.2586$

Item	p^*
Berger	0.3546
Chicken	0.2964
Fish	0.2478
Tofu	0.1011

Dice, no constraint

- Let $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$
- No other constraint
- Maximum entropy distribution is uniform distribution

$$p_i = 1/6$$

Dice, with constraint

- This example was used by Boltzmann
- Suppose n dices are thrown on the table
- The total number of spots showing is $n\alpha$
- What is the proportion of the dice are showing face i, $i = 1, 2, \ldots, 6$?

- Assume n_i dice show face i

- There are $\binom{n}{n_1,\ldots,n_6}$ possible configurations

- This is a macrostate indexed by $(n_1, \ldots, n_6)$ with $\binom{n}{n_1,\ldots,n_6}$ microstates, each having probability $1/6^n$.

- Constraint: $\sum_{i=1}^6 i n_i = n\alpha$.

- Using maximum entropy solution, we find

$$p_i^* = e^{\lambda i} / \sum_{i=1}^{6} e^{\lambda i}.$$

Maximum entropy classifier

- In some fields of machine learning, multinomial logic model is refer to as a maximum entropy classier
- Minimizes the amount of prior information built into the distribution
- X_i: feature vector, β_k: vector of weights for outcome k, Y_k: random outcome, $k = 1, \ldots, K$
- Takes the form

$$p(Y_i = k) = \frac{e^{\beta_k^\top X_i}}{1 + \sum_{\ell=1}^{K-1} e^{\beta_\ell^\top X_i}}, \quad k = 1, \ldots, K-1.$$

Maximum entropy spectrum estimation

- Given a stationary zero-mean stochastic process $\{X_i\}$

$$R(k) = EX_iX_{i+k}$$

- Goal: to learn the structure of the process, want to estimate $R(k)$ from samples of the process

- Challenge:
  estimate for low values of k are based on large number of samples
  estimate for high values of k are based on small number of samples

- Should we set low value $R(k)$ to 0?

- Burg: replace them with maximum entropy estimates

Summary

- Maximizing entropy minimizes the amount of prior information built into unknown distribution

- Maximum entropy distribution can be found explicitly

$$p_i^* = \frac{e^{\sum_{j=1}^m \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

- Maximum entropy principle widely used