Lecture 09: Density Estimation

Course Teacher: Md. Sharifful Islam Bhuyan

Fundamental Learning Tasks

- Variations of probabilistic view subsume most machine learning problems
  - Classification: conditional discrete distribution
  - Regression: conditional continuous distribution
  - Clustering: mixture model
  - Dimensionality reduction: latent factor model
  - Density estimation: Bayesian network representing joint distribution

Probabilistic Model

- Assume data generated from a probability distribution $x_1, \dots, x_N \sim p(\mathbf{d}|h)$
- Assume the observations to be independent & identically distributed (i.i.d)
- Learn/estimate parameter h
- Features/attributes/labels are now random variable (independent in NB)
- Learning joint distribution from data, other distribution can be deduced
- Benefits of probabilistic model
  - Represent uncertainties (Dice rolling: always 6)
  - imperfect knowledge (decision tree: Impure class)
  - Handle missing and noisy data in a principled way

Estimation Techniques

- Prior, likelihood, posterior, partition function/marginal
- Typically maximize log-likelihood (no effect due to monotonicity)

$$p(h_i|\mathbf{d}) = \frac{p(\mathbf{d}|h_i)p(h_i)}{p(\mathbf{d})}$$

$$h_{MLE} = \arg\max_{h_i} p(\mathbf{d}|h_i) = \arg\min_{h_i}(-\log p(\mathbf{d}|h_i))$$

$$h_{MAP} = \arg\max_{h_i} p(\mathbf{d}|h_i)p(h_i) = \arg\min_{h_i}(-\log p(\mathbf{d}|h_i) - \log p(h_i))$$

$$p_{Bayesian}(h_i|\mathbf{d}) = \frac{p(\mathbf{d}|h_i)p(h_i)}{p(\mathbf{d})} = \frac{p(\mathbf{d}|h_i)p(h_i)}{\sum_i p(\mathbf{d}|h_i)p(h_i)}$$