Machine Learning and Computational Statistics
Conditional Probability Models

Intro: This document consists of concepts and exercises related to Conditional Probability Model. The core idea lies in the frequentist's methodology in which parameters (e.g., θ) is not a random variable.

Learning objectives of Conditional Probability Models include:

1. Point estimation
   (a) One point of statistical problem is point estimation
   (b) A statistic s = s(D) is any function of the data
   (c) A statistic $\hat{\theta}$ (D) taking values in Θ is a point estimator of θ
   (d) A good point estimator will have $\hat{\theta} \approx \theta$
   (e) Desirable statistical properties of point estimators are:
      i. Consistency: As data size $n \to \inf$, we get $\hat{\theta}_n \to \theta$.
      ii. Efficiency: $\hat{\theta}_n$ is as accurate as we can get from a sample of size n.
   (f) Maximum likelihood estimators (MLE) are consistent and efficient under reasonable conditions.
2. Common approaches to getting a point estimator for θ
   (a) Maximum Likelihood (MLE)
      i. The likelihood of an estimate of a probability distribution for some data D.

$$p(D|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

$$L_D(\theta) = logp(D|\theta)$$

      ii. The MLE for some parameter θ of a probability model.

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmax}} L_D(\theta)$$

iii. Describe the basic structure of a linear probabilistic model, in terms of 1. a parameter θ of the probablistic model, 2. a linear score function, 3. a transfer function (kin to a "response function" or "inverse link" function, though we have relaxed requirements on the parameter theta).

iv. We can use MLE to find w by represent the parameter (e.g., θ) as a function of x, w and differentiate log-likelihood of the set-up probabilistic model to find w.

v. Common transfer functions for some distributions and why:

A. bernoulli

B. poisson: $\lambda = \psi(w^T x) = \exp^{w^T x}$

C. gaussian: $y|x \sim \mathcal{N}(w^T x, \sigma^2)$

D. categorical distributions: $\psi(s) = \psi(s_0, s_1, ...s_n) = (\frac{\exp_0^s}{\sum_{i=1}^n \exp^{s_i}}, \frac{\exp_1^s}{\sum_{i=1}^n \exp^{s_i}}, ..., \frac{\exp_n^s}{\sum_{i=1}^n \exp^{s_i}})$

(b) Method of Moment

(c) Maximum A Posterior (MAP) (Bayesian method)

$$P(\theta|D) = \frac{P(D|\theta) * P(\theta)}{P(D)} \propto P(D|\theta) * P(\theta)$$

$$\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\text{argmax}} P(D|\theta) * P(\theta) \\
&= \underset{\theta}{\text{argmax}} log P(D|\theta) * P(\theta) \\
&= \underset{\theta}{\text{argmax}} log \prod_{i=1}^{n} P(y_i|x_i, w) * P(\theta) \\
&= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} log P(y_i|x_i, w) * P(\theta)
\end{aligned}$$

3. MLE vs. MAP

4. Concept check questions Conditional Probability Models