Introduction to MLE and MAP

Bayes Rule for random variables

- Given that X and Θ are random variables,

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)}{f_X(x)}$$

- Based directly off from the Bayes theorem for sets and probability.

Basic Probability

- Probability triplet: $(\Omega, \mathcal{F}, P)$
  - Ω is a set
  - $\mathcal{F}$ is a set of some subsets of Ω
  - P is a function such that for $A \in \mathcal{F}$, $P(A) \in [0, 1]$
- Random Variable X is a function

$$X : \Omega \to \mathbb{R} \text{ (or something similar)}$$

and we denote $P(X = x)$ or $P_X(x)$ as

$$P_X(x) = P(\{\omega \in \Omega, X(\omega) = x\})$$

Conditional Probability

- Given $A, B \in \mathcal{F}$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- For random variables X and Y,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Likelihood

- Let $\mathbf{x} = x_1, \dots, x_n$ be our data set. We have a family of distribution that we guess our data comes from. Let θ be our estimate of the best fit parameter of our distribution.
- Likelihood gives the probability of θ matching the data $\mathbf{x}$

$$\mathcal{L}(\theta|\mathbf{x})$$

- Using Bayes theorem, we have

$$\mathcal{L}_{\Theta|X}(\theta|x) = f_{X|\Theta}(\mathbf{x}|\theta)\frac{f_\Theta(\theta)}{f_X(\mathbf{x})}$$

Maximum Likelihood Principle

- To estimate the parameter for the family of distributions, we take the mode of the distribution (or maximum)

$$\theta_{MLE} = \max_{\theta} \mathcal{L}(\theta|\mathbf{x})$$

- $f_X(\mathbf{x})$ is constant since it is $\int_\theta f_{X,\Theta}(x,\theta)d\theta$ and we integrate over all of θ.
- If we have no prior information about θ, then $f_\Theta(\theta)$ is constant (or uniformly distributed)
- Then,

$$\theta_{MLE} = \arg\max_{\theta} f(\mathbf{x}|\theta)$$

Example : Coin Flips

- Let θ be the probability of heads in a coin toss.
- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be the observations and let X be the random variable for the number of heads
- $f(\mathbf{x}|\theta)$ is Bernoulli and is given by

$$\binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

where k is the number of heads and n is the number of flips

Example : Coin Flips (2)

- To find the max we take

$$\frac{df}{d\theta} = k\theta^{k-1}(1-\theta)^{n-k} - (n-k)\theta^{k}(1-\theta)^{n-k-1}$$

- Setting the derivative to 0 we have

$$k(1-\theta) - (n-k)\theta = 0$$

- Solving for θ we have

$$\theta = \frac{k}{n}$$

- MLE estimate is total number of heads over total flips

Cross Entropy

- **Entropy** : Given a probability distribution P, entropy is given by (for the discrete case)

$$-\mathbb{E}[\log P] = -\sum_x P(x) \log P(x)$$

- **KL Divergence**: For probability distribution P and Q, it is the expectation of the log likelihood,

$$\sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right) = \sum_x \left(P(x) \log P(x) - P(x) \log Q(x) \right)$$

- **Cross Entropy**: Sum of entropy of P and KL divergence between P and Q,

$$-\sum_x P(x) \log Q(x)$$

Deep Learning (Multi-Class Classification)

- $\mathbf{x}$ is the data-set and θ is the weights of the neural network
- Data point x_i. Ground truth label c_i.
- Neural network outputs $\mathbf{s_i}$. s_{c_i} is the output of the neural network for class c
- Using the training data as P and the neural network output s Q, we define the cross entropy between P and Q.
- P is zero for all classes except the truth label
- Cross entropy formula which we use for likelihood

$$\mathcal{L}(\theta|\mathbf{x}) = -\sum_{i=1}^{n} \log(s_{c_i}^{\theta})$$

- Use backpropagation to find the minimum of the likelihood function and find the best weights of the neural network.

MAP

- Given that we have a prior distribution

$$\theta_{MAP} = \arg\max_{\theta} f_{X|\Theta}(\mathbf{x}|\theta) f_{\Theta}(\theta)$$

- Log version

$$\theta_{MAP} = \arg\max_{\theta} \left[\log f_{X|\Theta}(\mathbf{x}|\theta) + \log f_{\Theta}(\theta)\right]$$

- $f_{\Theta}(\theta)$ is another function we have to come up with.

Questions