Loss Functions Seminar 2

Old Dominion Vision Lab Seminar

Alex Glandon

2025

Review the Problem

- Point Estimation
  - Consider a probability distribution parameterized by θ
    - $X \sim f_x(x|\theta)$
  - We want to estimate θ
  - We consider an estimate $\hat{\theta}$ is good if the loss $l\left(\theta, \hat{\theta}\right)$ is small
  - Minimize Risk = Minimize Expected Loss
  - No general solution, as E[$l(\theta, \hat{\theta})$] depends on θ

Review the Problem

- Prediction
  - Consider a data distribution with examples X, perhaps images, and Y can take values of classes $y \in 1, \ldots, n$
  - $X, Y \sim f_{x,y}(x, y)$
  - This is a mixture model, because
  - We want to estimate $f_{y|x}(y|x) = \delta(y - i_x)$
  - In other words, assuming a ground truth label exists for each image, the conditional distribution has minimal entropy meaning it is delta function of the correct label
  - Our estimate is an approximate model of the conditional distribution parameterized by a set of weights, collectively called θ
  - $\hat{f}_{y|x,\hat{\theta}}(y|x, \hat{\theta})$
  - We could use KL divergence to measure the difference from $\hat{f}$ to the true distribution f
  - However, this would be a function of x, and we need a loss that we can compute for a finite sample of $x_1, x_2, \ldots, x_m$

KL divergence and cross entropy

- KL divergence from a predicted distribution Q to a true data distribution P is
- $D_{KL}(P||Q) = H(P,Q) - H(P)$
- If the cross entropy is the same, the KL divergence is higher when the true distribution has a low entropy
- Example 1:
- P is a true coin toss with 75% heads, 25% tails
- Q is a predicted coin with 50% heads, 50% tails
- $H(P,Q) =$
- $-P(true = heads) \cdot \log\big(P(pred = heads)\big) - P(true = tails) \cdot \log\big(P(pred = tails)\big) =$
- $-\frac{3}{4} \cdot \log\left(\frac{1}{2}\right) - \frac{1}{4} \cdot \log\left(\frac{1}{2}\right) = 1$
- $H(P) =$
- $-P(true = heads) \cdot \log\big(P(true = heads)\big) - P(true = tails) \cdot \log\big(P(true = tails)\big) =$
- $-\frac{3}{4} \cdot \log\left(\frac{3}{4}\right) - \frac{1}{4} \cdot \log\left(\frac{1}{4}\right) = 0.811$
- $D_{KL}(P||Q) = 1 - 0.811 = 0.189$

KL divergence and cross entropy

- KL divergence from a predicted distribution Q to a true data distribution P is
- $D_{KL}(P||Q) = H(P,Q) - H(P)$
- If the cross entropy is the same, the KL divergence is higher when the true distribution has a low entropy
- Example 2:
- P is a true coin toss with 50% heads, 50% tails
- Q is a predicted coin with 50% heads, 50% tails
- $H(P,Q) =$
- $-P(true = heads) \cdot \log\big(P(pred = heads)\big) - P(true = tails) \cdot \log\big(P(pred = tails)\big) =$
- $-\frac{1}{2} \cdot \log\left(\frac{1}{2}\right) - \frac{1}{2} \cdot \log\left(\frac{1}{2}\right) = 1$
- $H(P) =$
- $-P(true = heads) \cdot \log\big(P(true = heads)\big) - P(true = tails) \cdot \log\big(P(true = tails)\big) =$
- $-\frac{1}{2} \cdot \log\left(\frac{1}{2}\right) - \frac{1}{2} \cdot \log\left(\frac{1}{2}\right) = 1$
- $D_{KL}(P||Q) = 1 - 1 = 0$

KL divergence and cross entropy

- If we want to minimize KL divergence
- $argmin_{\hat{f}}\, D_{KL}(f||\hat{f}) = H(f,\hat{f}) - H(f)$
- We can minimize cross entropy
- $argmin_{\hat{f}}\, D_{KL}(f||\hat{f}) = argmin_{\hat{f}}\, H(f,\hat{f})$

Cross Entropy, ML, and MAP (Bayesian)

- I will explain from point estimation perspective
- The same applies to prediction, for example my first seminar on MAP for prediction

Point Estimation

- $argmin_{\hat{\theta}}\, l(\theta, \hat{\theta})$ has no general solution
- So we make a constraint on the problem
- UMVU estimation
- Equivariance
- Maximum likelihood (meaning we lose the ability to define a loss)
- Maximum a posterior (Bayesian)

Maximum Likelihood

- I didn’t explain this for point estimation in last seminar
- As an example, take normal distribution
- We know for normal the sample mean is unbiased, and the sample variance is biased
- Maximum likelihood of normal gives sample mean and sample variance as estimators of mean and variance, so in the case of maximum likelihood for variance, we can a problem with being biased

Maximum Likelihood Large Sample Size

- If $X_1, X_2, \dots, X_n$ are i.i.d.
- $l\left(\hat{\theta}\middle|x_i\right)$ is the likelihood given sample i
- $L\left(\hat{\theta}\middle|x_1, \dots, x_n\right) = \Pi_{i=1}^{n} l\left(\hat{\theta}\middle|x_i\right)$ is the product of likelihoods for i.i.d. samples
- Theorem: $P\left(L(\theta|x_1, \dots, x_n) > L\left(\hat{\theta}\middle|x_1, \dots, x_n\right)\right) \to 1$ as $n \to \infty$ for any $\hat{\theta} \neq \theta$
- Meaning if the prediction is not equal to the ground truth, the likelihood will be larger for the ground truth than for the prediction as that sample size gets large

Maximum Likelihood Large Sample Size

- Using concept of KL divergence again
- $D_{KL}(f_\theta || f_{\hat{\theta}}) = E[\log(\frac{f_\theta(X)}{f_{\hat{\theta}}(X)})]$ where $X \sim f_\theta(x)$
- Jensen's inequality says a strictly convex function of an integral is less than the integral of the convex function (negative log is strictly convex)
- $E\left[\log\left(\frac{f_\theta(X)}{f_{\hat{\theta}}(X)}\right)\right] = E\left[-\log\left(\frac{f_{\hat{\theta}}(X)}{f_\theta(X)}\right)\right] > -\log\left(\mathrm{E}\left[\frac{f_{\hat{\theta}}(X)}{f_\theta(X)}\right]\right)$
- $\mathrm{E}\left[\frac{f_{\hat{\theta}}(X)}{f_\theta(X)}\right] = \int_{-\infty}^{\infty} \frac{f_{\hat{\theta}}(x)}{f_\theta(x)} f_\theta(x)dx = \int_{-\infty}^{\infty} f_{\hat{\theta}}(x)dx = 1$
- $-\log\left(\mathrm{E}\left[\frac{f_{\hat{\theta}}(X)}{f_\theta(X)}\right]\right) = -\log(1) = 0$
- KL divergence is non-negative

Maximum Likelihood Large Sample Size

- Theorem: $P\left(L(\theta|x_1,\dots,x_n) > L\left(\hat{\theta}\middle|x_1,\dots,x_n\right)\right) \to 1$ as $n \to \infty$ for any $\hat{\theta} \neq \theta$
- Proof:
- From previous slide about KL divergence, for any $\hat{\theta} \neq \theta$
- $-\log\left(E\left[\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right]\right) = 0$
- From Jensen's inequality
- $-\log\left(E\left[\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right]\right) < \mathrm{E}\left[-\log\left(\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right)\right]$
- $0 < \mathrm{E}\left[-\log\left(\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right)\right]$
- $0 < -\mathrm{E}\left[\log\left(\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right)\right]$
- $\mathrm{E}\left[\log\left(\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right)\right] < 0$

Maximum Likelihood Large Sample Size

- $\mathrm{E}\left[\log\left(\frac{\iota(\hat{\theta}|X)}{\iota(\theta|X)}\right)\right] < 0$
- The law of large numbers (the sample average converges to the true expectation), using weak version
- Let the function of X_i given by the quotient inside the sum be a random variable below
- $\forall \epsilon_0 > 0, \lim_{n\to\infty} P\left(\left|\frac{1}{n}\Sigma_i \log\left(\frac{\iota(\hat{\theta}|X_i)}{\iota(\theta|X_i)}\right) - \mathrm{E}\left[\log\left(\frac{\iota(\hat{\theta}|X_i)}{\iota(\theta|X_i)}\right)\right]\right| < \epsilon_0\right) = 1$
- Implies
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{\iota(\hat{\theta}|X_i)}{\iota(\theta|X_i)}\right) < 0\right) = 1$
- Analysis on next 2 slides

Maximum Likelihood Large Sample Size

- $\forall \epsilon_0 > 0, \lim_{n\to\infty} P\left(\left|\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) - \mathrm{E}\left[\log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right)\right]\right| < \epsilon_0\right) = 1$
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) - \epsilon_0 < \mathrm{E}\left[\log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right)\right] < \frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) + \epsilon_0\right) = 1$
- If P(rain and clouds)=100%, then P(rain)=100%
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) - \epsilon_0 < \mathrm{E}\left[\log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right)\right]\right) = 1$

Maximum Likelihood Large Sample Size

- $\forall\epsilon_0 \lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) - \epsilon_0 < \mathrm{E}\left[\log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right)\right]\right) = 1$
- $\mathrm{E}\left[\log\left(\frac{l(\hat{\theta}|X)}{l(\theta|X)}\right)\right] < 0$ means $\exists\epsilon_1 > 0$
- $\mathrm{E}\left[\log\left(\frac{l(\hat{\theta}|X)}{l(\theta|X)}\right)\right] + \epsilon_1 = 0$
- $\exists\epsilon_0 < \epsilon_1 \rightarrow$
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) - \epsilon_0 < -\epsilon_1\right) = 1$
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) < -\epsilon_1 + \epsilon_0\right) = 1$
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) < \epsilon_2\right) = 1$, $\epsilon_2 < 0$
- Again, if P(rain and clouds)=100%, then P(rain)=100%
- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) < 0\right) = 1$

Maximum Likelihood Large Sample Size

- $\lim_{n\to\infty} P\left(\frac{1}{n}\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) < 0\right) = 1$
- $\lim_{n\to\infty} P\left(\Sigma_i \log\left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) < 0\right) = 1$
- $\lim_{n\to\infty} P\left(\log\left(\Pi_i \left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right)\right) < 0\right) = 1$
- $\lim_{n\to\infty} P\left(\Pi_i \left(\frac{l(\hat{\theta}|X_i)}{l(\theta|X_i)}\right) < 1\right) = 1$
- $\lim_{n\to\infty} P\left(\frac{\Pi_i\, l(\hat{\theta}|X_i)}{\Pi_i\, l(\theta|X_i)} < 1\right) = 1$
- $\lim_{n\to\infty} P\left(\Pi_i\, l(\theta|X_i) > \Pi_i\, l(\hat{\theta}|X_i)\right) = 1$
- $\lim_{n\to\infty} P\left(L(\theta|X_1, \dots, X_m) > L(\hat{\theta}|X_1, \dots, X_m)\right) = 1$

Maximum Likelihood

- Equivalent to minimizing mean squared loss for regression
- Equivalent to minimizing cross entropy loss for classification
- May not be the best if the number of samples is low
  - For example, if number of samples if low, we would not use maximum likelihood for normal variance estimation, we would use the unbiased estimator
- May not be the best if the loss function is customized based on the problem
  - For example, consider a loss that penalizes small error more than large errors. According to Berger decision theory text, the first million dollars is worth more than the second million dollars, is an example of a loss function that is more complicated

MAP estimation

- Bayesian estimation is a point estimation
- Even in maximum likelihood, we have a distribution, but we find the mode or the max of the distribution
- In Bayesian, we use a prior to weight the distribution, but again we need to select a parameter to use in prediction, so we take the mode or the max of the posterior
- Considering a distribution is not special to Bayesian, is it just probabilistic
  - We also use distributions for UMVU, maximum likelihood, equivariance, minimax
- Of course there are ways to use Bayesian other than point estimation

References

[1] Theory of Point Estimation - Lehmann and Casella