CSE 546: Machine Learning **Lecture 1**

Overview / Maximum Likelihood Estimation

Instructor: Sham Kakade

1 What is Machine Learning?

Machine learning is the study of algorithms which improve their performance with experience. The area combines ideas from both computer science and statistics (and numerous other areas) for the simple reason that statistics is the means by which we model the natural world and computer science is the study of algorithms, which are relevant for manipulating such models.

2 Maximum Likelihood Estimation

In many machine learning (and statistics) questions, we focus on estimating parameters of a model.

2.1 Estimating the bias of a coin

Let's start by estimating the bias of a coin.

Suppose the probability of heads is θ_* and the probability of tails is $1 - \theta_*$. The parameter θ_* is often referred to as the bias of the coin.

Suppose we observe some sequence of coin flips $\mathcal{S}$. Suppose the flips are identically and independently distributed (i.i.d.). The probability of observing a sequence of flips $\mathcal{S}$ is:

$$\Pr(\mathcal{S}|\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

where N_H and N_T are the number of heads and tails, respectively, in the sequence.

The *maximum likelihood estimator* is the parameter which maximizes this function:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \Pr(\mathcal{S}|\theta) = \arg\max_{\theta} \log \Pr(\mathcal{S}|\theta)$$

where the last step follows since the $\log$ function is monotonically increasing.

For this particular case, we have that:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log(\theta^{N_H}(1-\theta)^{N_T})$$

We can minimize by this function by finding the θ such that:

$$0 = \frac{\partial}{\partial\theta} \log(\theta^{N_H}(1-\theta)^{N_T}) = \frac{N_H}{\theta} - \frac{N_T}{1-\theta}$$

Doing this, we find that:

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$

Note that this estimator is *unbiased* in the following sense:

$$\mathbb{E}\hat{\theta} = \theta_*$$

where the expectation is over the sequence of coin flips.

2.2 Estimating a mean

Let us now consider the problem of estimating a mean. Suppose we have a distribution $\Pr(X)$, and we wish to estimate the mean of this distribution. In particular, we observe data $\mathcal{S} = x_1, x_2, \ldots x_N$. What is an estimate of the mean?

Now let us model the data under a normal distribution.

$$\Pr(X) \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Here, we have that the log likelihood function is:

$$\log \Pr(\mathcal{S}|\mu) = \sum_i \log \Pr(X = x_i) = -N \log \sqrt{2\pi\sigma^2} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Again, to find the maximum likelihood estimate, we simply find the maxima of this function. By doing this, we obtain that the maximum likelihood estimator is:

$$\hat{\mu} = \frac{1}{N} \sum_i x_i$$

Note that that this estimate does not depend on knowledge of σ.

2.3 Maximum Likelihood Estimation: an optimization problem

More generally, suppose we have a probability model of our data $\Pr(\mathcal{S}|\theta)$.

If we view $\Pr(\mathcal{S}|\theta)$ as a function of θ (keeping $\mathcal{S}$ fixed), then this is referred to the *likelihood function*. We can define the log likelihood function as:

$$L_{\mathcal{S}}(\theta) = \log \Pr(\mathcal{S}|\theta)$$

Importantly, note that the functional dependence of the likelihood function is on θ (and $\mathcal{S}$ is considered fix).

The *maximum likelihood estimator* is the parameter which maximizes this function:

$$\hat{\theta}_{MLE} = \arg\max L_{\mathcal{S}}(\theta)$$

When it is clear from context, we sometimes drop the MLE subscript, and refer to this estimate by $\hat{\theta}$.

Importantly, note that computing the MLE is an *optimization* problem. If we know that the θ lies in some set Θ, then we solve the problem:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L_{\mathcal{S}}(\theta)$$

which is a constrained optimization problem.

Why is this idea appealing? In general, we want an estimate $\hat{\theta}$ which is accurate in some quantifiable sense. Let us see make a crude (and asymptotic) justification of this idea.

Suppose our model posits that our data $\{z_1, \ldots z_N\}$ are i.i.d. And let us suppose that our data are in fact generated by some $\log \Pr(z|\theta_*)$ for some θ_* in our model class.

We have that:

$$\hat{\theta}_{MLE,N} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log \Pr(z_i|\theta)$$

Now for large enough N, we may hope that this function well approximates the expected log likelihood function:

$$\mathbb{E} \log \Pr(z|\theta)$$

where the expectation is over the random variable z.

Now let us consider the following optimization problem:

$$\arg\max_{\theta} \mathbb{E} \log \Pr(z|\theta)$$

Under our assumption that the data are generating according to θ_*, one can show that θ_* is a maximizer of the above (through a convexity argument). To see this, note that since the $\log$ is a concave function, we have that:

$$\begin{aligned}
\mathbb{E} \log \Pr(z|\theta_*) - \mathbb{E} \log \Pr(z|\theta) &= - \mathbb{E} \log \frac{\Pr(z|\theta)}{\Pr(z|\theta_*)} \\
&\geq - \log \mathbb{E} \frac{\Pr(z|\theta)}{\Pr(z|\theta_*)} \\
&= - \log \sum_z \Pr(z|\theta_*) \frac{\Pr(z|\theta)}{\Pr(z|\theta_*)} \\
&= - \log \sum_z \Pr(z|\theta) \\
&= - \log 1 \\
&= 0
\end{aligned}$$

The

Furthermore, under mild regularity conditions, for sufficiently large N, we have that $\hat{\theta}_{MLE,N}$ converges to θ_*, i.e. it is *consistent*. In particular, any estimation procedure (one which provides an estimate $\hat{\theta}_N$ given N samples of the data) is said to be *consistent* if $\hat{\theta}_N$ converges to the θ_*. Furthermore, in many cases, the MLE is statistically *efficient*. In particular, in many cases, it is the case that, in the limit of large enough N, no other (unbiased) estimator has lower variance.

3 How good are these estimates?

Now what are the quality of these estimates? Note that even if our Gaussian assumption is not correct (sometimes referred to as *model misspecification*, we might still expect that our estimates be reasonable.

3.1 Review: the central limit theorem

Theorem 3.1. *Suppose $X_1, X_2, \ldots X_n$ is a sequence of independent, identically distributed (i.i.d.) random variables with mean μ and variance σ_2. Let $\bar{X}_n = \sum_{i=1}^{n} X_i$. Under certain mild conditions (in particular, suppose that the* *moment generating function $M_X(\lambda)$ exists for all λ in a neighborhood of 0), we what that for all z,*

$$\lim_{n\to\infty} \Pr\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z\right) = \Phi(z)$$

where $\Phi(\cdot)$ is the standard normal CDF.

Roughly speaking, this says that, in the limit of large n, $\bar{X}_n$ is distributed according to a Gaussian with mean μ and variance σ^2/n

3.2 Probability Approximately Correct (PAC) statements

Theorem 3.2. *(Chernoff-Hoeffding Bound) Let $X_1, X_2, \ldots X_N$ be N i.i.d. random variables with $X_i \in [a, b]$ (with probability one). Then for all $\epsilon > 0$ we have:*

$$\Pr(\frac{1}{N}\sum_{i=1}^{m} X_i - \mathbb{E}[X] > \epsilon) \le e^{-\frac{2N\epsilon^2}{(b-a)^2}}$$

Suppose that for the mean estimation case considered earlier, that our random variable X is bounded in $[0, 1]$ with probability 1. Then we have that with probability greater than $1 - \delta$,

$$|\hat{\theta} - \mathbb{E}[X]| \le \sqrt{\log/2\delta 2N}$$

To see this note that:

$$\Pr(|\hat{\theta} - \mathbb{E}[X]| \ge \epsilon) \le \Pr(\frac{1}{N}\sum X_i - \mathbb{E}[X] > \epsilon) + \Pr(\frac{1}{N}\sum X_i - \mathbb{E}[X] < -\epsilon) \le 2e^{-\frac{2N\epsilon^2}{(b-a)^2}}$$

And choosing $\epsilon = \sqrt{\log/2\delta 2N}$ completes the argument.

3.3 Bernstein bound

With stronger *concentration* bounds (such as the Bernstein bound), one can show that with probability greater than $1 - \delta$,

$$\frac{1}{N}\sum_{i=1}^{m} X_i - \mathbb{E}[X] \le \sqrt{\frac{2\text{Var}(X)\log 1/\delta}{N}} + \frac{2B\log(1/\delta)}{N}$$

where Var(X) is the variance of X and B is an upper bound on X.

4 The Basic Idea

- collect some data...
- choose a hypothesis class or model
- choose a loss function (the log loss was what we used in this lecture)
- choose an optimization method to minimize the loss