ECE 901
Lecture 13: Maximum Likelihood Estimation

R. Nowak

5/17/2009

The focus of this lecture is to consider another approach to learning based on maximum likelihood estimation. Unlike earlier approaches considered here we are willing to make somewhat stronger assumptions about the relation between features and labels. These are quite reasonable in many settings, in particular in many imaging applications.

Consider the classical signal plus noise model:

$$Y_i = f\left(\frac{i}{n}\right) + W_i, i = 1, \ldots, n$$

where W_i are iid zero-mean noises. Furthermore, assume that W_i have a distribution characterized by a *probability density function* (p.d.f.) $p(w)$ for some known density $p(w)$. Then

$$Y_i \sim p\left(y - f\left(\frac{i}{n}\right)\right) \equiv p_{f_i}(y)$$

since $Y_i - f\left(\frac{i}{n}\right) = W_i$.

In a setting like this it is quite common to consider the maximum likelihood approach - seek a "most-probable" explanation for the observations. Define the *likelihood* of the data to be the the p.d.f. of the observations $(Y_1, \ldots, Y_n)$

$$\prod_{i=1}^{n} p_{f_i}(Y_i) \ .$$

The maximum likelihood estimator seeks the model that maximizes the likelihood, or equivalently minimizes the negative log-likelihood

$$\sum i = 1^n - \log p_{f_i}(Y_i) \ . \tag{1}$$

We immediately notice the similarity between the empirical risk we had seen before and the negative log-likelihood. We will see that we can regard maximum likelihood estimation as our familiar minimal empirical risk when the loss function is chosen appropriately. In the meantime note that minimizing (1) yields our familiar square-error loss if W_i's are Gaussian. If the W_i's are Laplacian ($p_W(w) \propto e^{-c|w|}$) we get the sum of absolute errors. We can also consider non-additive models like the Poisson model (used often in medical imaging applications, like PET imaging)

$$Y_i \sim p(y|f(i/n)) = e^{-f(i/n)} \frac{f^y(i/n)}{y} \ ,$$

that gives rise to the following negative log-likelihood

$$-\log P\left(Y_i|f\left(i/n\right)\right) \ = \ f\left(i/n\right) - Y_i \log\left(f\left(i/n\right)\right) + \text{constant} \ ,$$

which is a very different loss function, but quite appropriate for many imaging problems.

1 Maximum Likelihood Estimation

Before we investigate maximum likelihood estimation for model selection, let's review some of the basis concepts. Let Θ denote a parameter space (e.g., $\Theta = R$, or $\Theta = \{\text{smooth functions}\}$). Assume we have observations

$$Y_i \overset{iid}{\sim} p_{\theta^*}(y), \quad i = 1, \ldots, n$$

where $\theta^* \in \Theta$ is a parameter determining the density of the $\{Y_i\}$. The ML estimator of θ^* is

$$\begin{aligned}
\hat{\theta}_n &= \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(Y_i) \\
&= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_\theta(Y_i) \\
&= \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(Y_i) \ .
\end{aligned}$$

Note that by the strong law of large numbers

$$\frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(Y_i) \overset{a.s.}{\to} E[-\log p_\theta(Y)] \ .$$

So we can use the negative log-likelihood as a proxy for $E[-\log p_\theta(Y)]$. Let's see why this is the think to do.

$$\begin{aligned}
E[\log p_{\theta^*}(Y) - \log p_\theta(Y)] &= E\left[\log \frac{p_{\theta^*}(Y)}{p_\theta(Y)}\right] \\
&= \int \log \frac{p_{\theta^*}(y)}{p_\theta(y)} p_{\theta^*}(y) dy \\
&\equiv K(p_\theta, p_{\theta^*}) \quad \text{the KL divergence} \\
&\geq 0 \quad \text{with equality iff} \, p_{\theta^*} = p_\theta.
\end{aligned}$$

Where K is the Kullback-Leibler divergence between to densities. This is a measure of the distinguishability between two different random variables. It is not a symmetric function so the order of the arguments is important. Furthermore it is always positive, and zero only if the two densities are identical

$$\begin{aligned}
-E\left[\log \frac{p_{\theta^*}(y)}{p_\theta(y)}\right] &= E\left[\log \frac{p_\theta(y)}{p_{\theta^*}(y)}\right] \\
&\leq \log E\left[\frac{p_\theta(y)}{p_{\theta^*}(y)}\right] \\
&= \log \int p_\theta(y) dy = 0 \\
&\Rightarrow \quad K(p_\theta, p_{\theta^*}) \geq 0
\end{aligned}$$

By showing that $E[\log p_{\theta^*}(Y) - \log p_\theta(Y)] \geq 0$ we immediately see that minimizing $E[-\log p_{\theta^*}(Y)]$ with respect to θ gets us close to the true model θ^*, exactly what we want to do.

1.1 Likelihood as a Loss Function

We can restate the maximum likelihood estimator in the general terms we are using in this course. We have n i.i.d observations drawn from an unknown distribution

$$Y_i \overset{i.i.d.}{\sim} p_{\theta^*} \quad , \ i = \{1, \ldots, n\}$$

where $\theta^* \in \Theta$. We can view p_{θ^*} as a member of a parametric class of distributions, $\mathcal{P} = \{p_\theta\}_{\theta\in\Theta}$. Our goal is to use the observations $\{Y_i\}$ to *select* an appropriate distribution (e.g., model) from $\mathcal{P}$. We would like the selected distribution to be close to p^*_θ in some sense.

We use the negative log-likelihood *loss function*, defined as $l(\theta, Y_i) = -\log p_\theta(Y_i)$. The **empirical risk** is

$$\hat{R}_n = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(Y_i).$$

We select the distribution that minimizes the empirical risk

$$\min_{p\in\mathcal{P}} -\sum_{i=1}^{n} \log p(Y_i) \;=\; \min_{\theta\in\Theta} -\sum_{i=1}^{n} \log p_\theta(Y_i)$$

In other words, the distribution we select is $\hat{p} := p_{\hat{\theta}_n}$, where

$$\hat{\theta}_n \;=\; \arg\min_{\theta\in\Theta} \quad -\sum_{i=1}^{n} \log p_\theta(Y_i)$$

The **risk** is defined as

$$R(\theta) = E[l(\theta, Y)] = -E[\log p_\theta(Y)].$$

And, the **excess risk** of θ is defined as

$$R(\theta) - R(\theta^*) = \int \log \frac{p_{\theta^*}(y)}{p_\theta(y)}\, p_{\theta^*}(y)\, \mathrm{d}y \equiv K(p_\theta, p_{\theta^*})\,.$$

We recognized that the excess risk corresponding to this loss function is simply the *Kullback-Leibler (KL) Divergence* or *Relative Entropy*, denoted by $K(p_{\theta_1}, p_{\theta_2})$. It is easy to see that $K(p_{\theta_1}, p_{\theta_2})$ is always non-negative and is zero if and only if $p_{\theta_1} = p_{\theta_2}$. This shows that θ^* minimizes the risk. The KL divergence measures how different two probability distributions are and therefore is natural to measure convergence of the maximum likelihood procedures.

1.2 Convergence of Log-Likelihood to KL Divergence

Since $\hat{\theta}_n$ maximizes the likelihood over $\theta \in \Theta$, we have

$$\sum_{i=1}^{n} \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} = \sum_{i=1}^{n} \log p_{\theta^*}(Y_i) - \log p_{\hat{\theta}_n}(Y_i) \le 0$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} - K(p_{\hat{\theta}_n}, p_{\theta^*}) + K(p_{\hat{\theta}_n}, p_{\theta^*}) \le 0$$

or re-arranging

$$K(p_{\hat{\theta}_n}, p_{\theta^*}) \le \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} - K(p_{\hat{\theta}_n}, p_{\theta^*}) \right|$$

Notice that the quantity

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta^*}(Y_i)}{p_\theta(Y_i)}$$

is an empirical average whose mean is $K(p_\theta, p_{\theta^*})$. By the law of large numbers, for each $\theta \in \Theta$,

$$\left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta^*}(Y_i)}{p_\theta(Y_i)} - K(p_\theta, p_{\theta^*}) \right| \overset{a.s.}{\to} 0$$

If this also holds for the sequence $\{\hat{\theta}_n\}$, then we have

$$K(p_{\hat{\theta}_n}, p_{\theta^*}) \leq \left| \frac{1}{n} \sum \log \frac{p_{\theta^*}(Y_i)}{p_{\hat{\theta}_n}(Y_i)} - K(p_{\hat{\theta}_n}, p_{\theta^*}) \right| \to 0 \, \text{as} \, n \to \infty$$

which implies that

$$p_{\hat{\theta}_n} \to p_{\theta^*}$$

which often implies that

$$\hat{\theta}_n \to \theta^*$$

in some appropriate sense (e.g., point-wise or in norm).

Example 1. *Gaussian Distributions*

$$p_{\theta^*}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta^*)^2}{2\sigma^2}}$$

$$\Theta = \mathbb{R}, \quad \{Y_i\}_{i=1}^n \overset{iid}{\sim} p_{\theta^*}(y)$$

$$\begin{aligned}
K(p_\theta, p_{\theta^*}) &= \int \log \frac{p_{\theta^*}(y)}{p_\theta(y)} p_{\theta^*}(y) dy \\
&= \frac{1}{2\sigma^2} \int [(y-\theta)^2 - (y-\theta^*)^2] p_{\theta^*}(y) dy \\
&= \frac{1}{2\sigma^2} E_{\theta^*}[(Y-\theta)^2] - \frac{1}{2\sigma^2} E_{\theta^*}[(Y-\theta^*)^2] \\
&= \frac{1}{2\sigma^2} E_{\theta^*}[-2Y(\theta+\theta^*) + \theta^2 + \theta^{*2}] \\
&= \frac{1}{2\sigma^2} \left[-2(\theta+\theta^*) E_{\theta^*}[Y] + \theta^2 + \theta^{*2} \right] \\
&= \frac{(\theta^*-\theta)^2}{2\sigma^2} \, .
\end{aligned}$$

$$\Rightarrow \theta^* \;\; \textit{maximizes} \;\; E[\log p_\theta(Y)] \;\; \textit{wrt} \;\; \theta \in \Theta$$

$$\begin{aligned}
\hat{\theta}_n &= \arg\max_\theta \{ -\sum (Y_i - \theta)^2 \} \\
&= \arg\min_\theta \{ \sum (Y_i - \theta)^2 \} \\
&= \frac{1}{n} \sum_{i=1}^n Y_i
\end{aligned}$$

1.3 Hellinger Distance

The KL divergence is not a distance function.

$$K(p_{\theta_1}, p_{\theta_2}) \neq K(p_{\theta_2}, p_{\theta_1})$$

Therefore, it is often more convenient to work with the Hellinger metric,

$$H(p_{\theta_1}, p_{\theta_2}) = \left(\int \left(p_{\theta_1}^{\frac{1}{2}} - p_{\theta_2}^{\frac{1}{2}} \right)^2 dy \right)^{\frac{1}{2}}$$

The Hellinger metric is symmetric, non-negative and

$$H(p_{\theta_1}, p_{\theta_2}) = H(p_{\theta_2}, p_{\theta_1})$$

and therefore it is a distance measure. Furthermore, the squared Hellinger distance lower bounds the KL divergence, so convergence in KL divergence implies convergence of the Hellinger distance.

Proposition 1.

$$H^2(p_{\theta_1}, p_{\theta_2}) \leq K(p_{\theta_1}, p_{\theta_2})$$

Proof:

$$\begin{aligned}
H^2(p_{\theta_1}, p_{\theta_2}) &= \int \left(\sqrt{p_{\theta_1}(y)} - \sqrt{p_{\theta_2}(y)}\right)^2 dy \\
&= \int p_{\theta_1}(y)dy + \int p_{\theta_2}(y)dy - 2\int \sqrt{p_{\theta_1}(y)}\sqrt{p_{\theta_2}(y)}dy \\
&= 2 - 2\int \sqrt{p_{\theta_1}(y)}\sqrt{p_{\theta_2}(y)}dy, \quad \text{since} \int p_\theta(y)dy = 1 \,\forall\theta \\
&= 2\left(1 - E_{\theta_2}\left[\sqrt{p_{\theta_1}(Y)/p_{\theta_2}(Y)}\right]\right) \\
&\leq 2\log\left(E_{\theta_2}\left[\sqrt{p_{\theta_2}(Y)/p_{\theta_1}(Y)}\right]\right), \quad \text{since} \, 1 - x \leq -\log x \\
&\leq 2E_{\theta_2}\left[\log\sqrt{p_{\theta_2}(Y)/p_{\theta_1}(Y)}\right], \quad \text{by Jensen's inequality} \\
&= E_{\theta_2}\left[\log(p_{\theta_2}(Y)/p_{\theta_1}(Y))\right] \equiv K(p_{\theta_1}, p_{\theta_2})
\end{aligned}$$

■

Note that in the proof we also showed that

$$H(p_{\theta_1}, p_{\theta_2}) = 2\left(1 - \int \sqrt{p_{\theta_1}(y)}\sqrt{p_{\theta_2}(y)}dy\right)$$

and using the fact $\log x \leq x - 1$ again, we have

$$H(p_{\theta_1}, p_{\theta_2}) \leq -2\log\left(\int \sqrt{p_{\theta_1}(y)}\sqrt{p_{\theta_2}(y)}dy\right)$$

The quantity inside the log is called the *affinity* between p_{θ_1} and p_{θ_2}:

$$A(p_{\theta_1}, p_{\theta_2}) = \int \sqrt{p_{\theta_1}(y)}\sqrt{p_{\theta_2}(y)}dy$$

This is another measure of closeness between p_{θ_1} and p_{θ_2}.

Example 2. *Gaussian Distributions*

$$p_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$$

$$
\begin{aligned}
& -2\log\int\sqrt{p_{\theta_1}(y)}\sqrt{p_{\theta_2}(y)}dy \\
= \;& -2\log\int\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\theta_1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\theta_2)^2}{2\sigma^2}}\right)^{\frac{1}{2}}dy \\
= \;& -2\log\left(\int\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\left[\frac{(y-\theta_1)^2}{4\sigma^2}+\frac{(y-\theta_2)^2}{4\sigma^2}\right]}dy\right) \\
= \;& -2\log\left(\int\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\left[\frac{\left(y-(\frac{\theta_1+\theta_2}{2})\right)^2}{2\sigma^2}+\frac{(\theta_1-\theta_2)^2}{8\sigma^2}\right]}dy\right) \\
= \;& -2\log e^{-\frac{(\theta_1-\theta_2)^2}{8\sigma^2}} \\
= \;& \frac{(\theta_1-\theta_2)^2}{4\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \quad & -2\log A(p_{\theta_1},p_{\theta_2}) = \frac{(\theta_1-\theta_2)^2}{4\sigma^2} \quad \textit{for Gaussian distributions} \\
\Rightarrow \quad & H^2(p_{\theta_1},p_{\theta_2}) \le \frac{(\theta_1-\theta_2)^2}{4\sigma^2} \quad \textit{for Gaussian.}
\end{aligned}
$$

Summary

$Y_i \overset{iid}{\sim} p_{\theta^*}$

1. Maximum likelihood estimator maximizes the empirical average

$$\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(Y_i)$$

(our empirical risk is negative log-likelihood)

2. θ^* maximizes the expectation

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(Y_i)\right]$$

(the risk is the expected negative log-likelihood)

3.

$$\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(Y_i) \overset{a.s.}{\to} E\left[\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(Y_i)\right]$$

so we expect some sort of concentration of measure.

4. In particular, since

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\theta^*}(Y_i)}{p_\theta(Y_i)} \overset{a.s.}{\to} K(p_\theta,p_{\theta^*})$$

we might expect that $K(p_{\hat{\theta}_n},p_{\theta^*})\to 0$ for the sequence of estimates $\{p_{\hat{\theta}_n}\}_{n=1}^{\infty}$.

So, the point is that maximum likelihood estimator is just a special case of a loss function in learning. Due to its special structure, we are naturally led to consider KL divergences, Hellinger distances, and Affinities.