ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** Matt Bayer

Lecture 7: Hypothesis Testing and KL Divergence

1 Introducing the KL Divergence

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} q(x)$ and we have two models for $q(x)$, $p_0(x)$ and $p_1(x)$. The likelihood ratio is

$$\Lambda = \prod_{i=1}^{n} \frac{p_1(x_i)}{p_0(x_i)}$$

The log likelihood ratio, normalized by dividing by n, is then

$$\hat{\Lambda}_n = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_1(x_i)}{p_0(x_i)}$$

Note that $\hat{\Lambda}_n$ is itself a random variable, and is in fact a sum of iid random variables $L_i = \log \frac{p_1(x_i)}{p_0(x_i)}$ which are independent because the x_i are. In addition, we know from the strong law of large numbers that for large n,

$$\begin{aligned}
\hat{\Lambda}_n &\overset{a.s.}{\to} \mathbb{E}\left[\hat{\Lambda}_n\right] \\
\mathbb{E}\left[\hat{\Lambda}\right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[L_i\right] \\
&= \mathbb{E}\left[L_1\right] \\
&= \int \log \frac{p_1(x)}{p_0(x)} q(x) dx \\
&= \int \log \left(\frac{p_1(x)}{p_0(x)} \frac{q(x)}{q(x)} \right) q(x) dx \\
&= \int \left[\log \frac{q(x)}{p_0(x)} - \log \frac{q(x)}{p_1(x)} \right] q(x) dx \\
&= \int \log \frac{q(x)}{p_0(x)} q(x) dx - \int \log \frac{q(x)}{p_1(x)} q(x) dx
\end{aligned}$$

The quantity $\int \log \frac{q(x)}{p(x)} q(x) dx$ is known as the *Kullback-Leibler Divergence* of p from q, or the *KL divergence* for short. We use the notation

$$D(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

for continuous random variables, and

$$D(q||p) = \sum_i q_i \log \frac{q_i}{p_i}$$

for discrete random variables. The above expression for $\mathbb{E}\left[\hat{\Lambda}_n\right]$ can then be written as

$$\mathbb{E}\left[\hat{\Lambda}_n\right] = D(q||p_0) - D(q||p_1)$$

Therefore, for large n, the log likelihood ratio test (LRT) $\mathbb{E}\left[\hat{\Lambda}_n\right] \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$ is approximately performing the comparison

$$D(q||p_0) - D(q||p_1) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

In particular, for $\lambda = 0$–i.e., if the probabilities for the two models are equal at the threshold value–we have the test

$$D(q||p_0) \underset{H_0}{\overset{H_1}{\gtrless}} D(q||p_1)$$

For this case, using the LRT is equivalent to selecting the model that is ”closer” to q in the sense of KL divergence.

Example 1 *Suppose we have the hypotheses*

$$\begin{aligned} H_0 &: X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2) \\ H_1 &: X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2) \end{aligned}$$

Then we can calculate the KL divergence:

$$\begin{aligned} \log \frac{p_1(x)}{p_0(x)} &= \log \left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_1)^2\right]}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_0)^2\right]} \right) \\ &= -\frac{1}{2\sigma^2}\left[(x-\mu_1)^2 - (x-\mu_0)^2\right] \\ &= -\frac{1}{2\sigma^2}\left[-2x\mu_1 + \mu_1^2 + 2x\mu_0 - \mu_0^2\right] \\ D(p_1||p_0) &= \int \log p_1(x) \frac{p_1(x)}{p_0(x)} dx \\ &= \mathbb{E}_{p_1}\left[\log \frac{p_1}{p_0}\right] \\ &= \mathbb{E}_{p_1}\left[-\frac{1}{2\sigma^2}\left(-2x\mu_1 + \mu_1^2 + 2x\mu_0 - \mu_0^2\right)\right] \\ &= -\frac{1}{2\sigma^2}\left(2(\mu_0 - \mu_1)\mathbb{E}_{p_1}[x] + \mu_1^2 - \mu_0^2\right) \\ &= -\frac{1}{2\sigma^2}\left(-2mu_1^2 + \mu_1^2 + 2\mu_1\mu_0 - \mu_0^2\right) \\ &= \frac{1}{2\sigma^2}\left(\mu_0^2 - 2\mu_0\mu_1 + \mu_1^2\right) \\ &= \frac{(\mu_1 - \mu_0)^2}{2\sigma^2} \end{aligned}$$

So the KL divergence between two Gaussian distributions with different means and the same variance is just proportional to the squared distance between the two means. In this case, we can see by symmetry that $D(p_1||p_0) = D(p_0||p_1)$, *but in general this is not true.*

2 A Key Property

The key property in question is that $D(q||p) \geq 0$, with equality if and only if $q = p$. To prove this, we will need a result in probability known as Jensen's Inequality:

> **Jensen's Inequality:** If a function $f(x)$ is convex, then
>
> $$\mathbb{E}\left[f(x)\right] \geq f(\mathbb{E}\left[x\right])$$
>
> A function is *convex* if $\forall\ \lambda \in [0, 1]$
>
> $$f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda)f(y)$$
>
> The left hand side of this inequality is the function value at some point between x and y, and the right hand side is the value of a straight line connecting the points $(x, f(x))$ and $(y, f(y))$. In other words, for a convex function the function value between two points is always lower than the straight line between those points.

Now if we rearrange the KL divergence formula,

$$\begin{aligned} D(q||p) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= \mathbb{E}_q \left[\log \frac{q(x)}{p(x)}\right] \\ &= -\mathbb{E}_q \left[\log \frac{p(x)}{q(x)}\right] \end{aligned}$$

we can use Jensen's inequality, since $-\log z$ is a convex function.

$$\begin{aligned} &\geq -\log\left(\mathbb{E}_q\left[\frac{p(x)}{q(x)}\right]\right) \\ &= -\log\left(\int q(x)\frac{p(x)}{q(x)}dx\right) \\ &= -\log\left(\int p(x)dx\right) \\ &= -\log(1) \\ &= 0 \end{aligned}$$

Therefore $D(q||p) \geq 0$.

3 Bounding the Error Probabilities

The KL divergence also provides a means to bound the error probabilities for a hypothesis test. For this we will need to recall Hueffding's Inequality.

> **Hueffding's Inequality:** If $Z_1, \ldots, Z_n$ are iid and $a \leq Z_i \leq b$, $\forall\ i$, then

$$\mathbb{P}\left(\frac{1}{n}\sum_i Z_i - \mathbb{E}[Z] > \epsilon\right) \le e^{-2n\epsilon^2/c^2}$$

and

$$\mathbb{P}\left(\mathbb{E}[Z] - \frac{1}{n}\sum_i Z_i > \epsilon\right) \le e^{-2n\epsilon^2/c^2}$$

where $c^2 = (b-a)^2$.

Now suppose that p_0 and p_1 have the same support and that over that support they are both bounded away from zero and from above; i.e. $0 < \alpha \le p_i(x) \le \beta < \infty$, $i = 0, 1$. It then follows that

$$\log\frac{\alpha}{\beta} \le \log\frac{p_1(x_i)}{p_0(x_i)} \le \log\frac{\beta}{\alpha}$$

The quantity $\log\frac{p_1(x_i)}{p_0(x_i)}$ is just the random variable L_i. Thus L_i is bounded, and $\hat{\Lambda}_n$ is a sum of iid bounded random variables. This allows us to use Hueffding's Inequality. Now, consider the hypothesis test $\hat{\Lambda}_n \overset{H_1}{\underset{H_0}{\gtrless}} 0$. We can write the probability of false alarm as

$$\begin{aligned} P_{FA} &= \mathbb{P}\left(\hat{\Lambda}_n > 0|H_0\right) \\ &= \mathbb{P}\left(\hat{\Lambda}_n - \mathbb{E}\left[\hat{\Lambda}_n|H_0\right] > -\mathbb{E}\left[\hat{\Lambda}_n|H_0\right] \mid H_0\right) \end{aligned}$$

The quantity $-\mathbb{E}\left[\hat{\Lambda}_n|H_0\right]$ will be the ϵ in Hueffding's inequality. We can re-express it as

$$\begin{aligned} \mathbb{E}_{p_0}\left[\hat{\Lambda}_n|H_0\right] &= \int p_0(x)\log\frac{p_1(x)}{p_0(x)}dx \\ &= -\int p_0(x)\log\frac{p_0(x)}{p_1(x)}dx \\ &= -D(p_0||p_1) \end{aligned}$$

Finally applying Hueffding's inequality, we get

$$\begin{aligned} P_{FA} &= \mathbb{P}\left(\hat{\Lambda}_n - (-D(p_0||p_1)) > D(p_0||p_1) \mid H_0\right) \\ &\le e^{-2nD^2(p_0||p_1)/c^2} \end{aligned}$$

with $c^2 = \left(\log\frac{\beta}{\alpha} - \log\frac{\alpha}{\beta}\right)$.

Thus the probability of false alarm error is bounded by the KL divergence $D(p_0||p_1)$. As n or $D(p_0||p_1)$ increase, the error decreases exponentially. The bound for the probability of miss, the other type of error, can be found in a similar fashion:

$$\begin{aligned}
P_{FA} &= \mathbb{P}\left(\hat{\Lambda}_n < 0 \mid H_1\right) \\
&= \mathbb{P}\left(\hat{\Lambda}_n - D(p_1||p_0) < -D(p_1||p_0) \mid H_1\right) \\
&= \mathbb{P}\left(D(p_1||p_0) - \hat{\Lambda}_n > D(p_1||p_0) \mid H_1\right) \\
&\leq e^{-2nD^2(p_1||p_0)/c^2}
\end{aligned}$$