ECE 830 Spring 2015 Statistical Signal Processing

instructor: R. Nowak

Lecture 7: Hypothesis Testing and KL Divergence

1 Introducing the Kullback-Leibler 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 $p_1(x)$. In past lectures we have seen that the likelihood ratio test (LRT) is optimal, assuming that q is p_0 or p_1. The error probabilities can be computed numerically in many cases. The error probabilities converge to 0 as the number of samples n grows, but numerical calculations do not always yield insight into rate of convergence. In this lecture we will see that the rate is exponential in n and parameterized the Kullback-Leibler (KL) divergence, which quantifies the differences between the distributions p_0 and p_1. Our analysis will also give insight into the performance of the LRT when q is neither p_0 nor p_1. This is important since in practice p_0 and p_1 may be imperfect models for reality, q in this context. The LRT acts as one would expect in such cases, it picks the model that is closest (in the sense of KL divergence) to q.

To begin our discusion, recall 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 $\hat{\Lambda}_n \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$ is approximately performing the comparison

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

since $\hat{\Lambda}_n$ will be close to its mean when n is large. Recall that the minimum probability of error test (assuming equal prior probabilities for the two hypotheses) is obtained by setting $\lambda = 0$. In this case, 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 selecting the model that is "closer" to q in the sense of KL divergence.

Example 1 *Suppose we have the hypotheses*

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

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

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

A function is *convex* if $\forall\ \lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

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 the following tail boud for averages of independent subGaussian random variables.

> **SubGaussian Tail Bound:** If $Z_1, \ldots, Z_n$ are independent and $\mathbb{P}(|Z_i - \mathbb{E}Z_i| \geq t) \leq ae^{-bt^2/2}$, $\forall\ i$, then
>
> $$\mathbb{P}\left(\frac{1}{n} \sum_i Z_i - \mathbb{E}\left[Z\right] > \epsilon \right) \leq e^{-cn\epsilon^2}$$
>
> and
>
> $$\mathbb{P}\left(\mathbb{E}\left[Z\right] - \frac{1}{n} \sum_i Z_i > \epsilon \right) \leq e^{-cn\epsilon^2}$$
>
> with $c = \frac{b}{16a}$.
>
> **Proof:** Follows immediately from Theorems 2 and 3 in `http://nowak.ece.wisc.edu/ece901_concentration.pdf`.

Now suppose that p_0 and p_1 have the same support and that the log likelihood ratio statistic $L_i := \log \frac{p_1(x_i)}{p_0(x_i)}$ has a subGaussian distribution; i.e., $\mathbb{P}(|L_i - \mathbb{E}L_i| \geq t) \leq ae^{-bt^2/2}$. For example, if p_0 and p_1 are Gaussian distributions with a common variance, then Z_i is a linear function of x_i and thus is Gaussian (and hence subGaussian). Note that $\hat{\Lambda}_n = \frac{1}{n} \sum_i L_i$ is an average of iid subGaussian random variables. This allows us to use the tail bound above.

Consider the hypothesis test $\hat{\Lambda}_n \underset{H_0}{\overset{H_1}{\gtrless}} 0$. We will now assume that the data $X_1, \ldots, X_n \overset{iid}{\sim} q$, with q either p_0 or p_1. We can write the probability of false positive error as

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

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

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

Applying the tail bound, we get

$$\begin{aligned} P_{FP} &= \mathbb{P}\left(\hat{\Lambda}_n - (-D(p_0||p_1)) > D(p_0||p_1) \mid H_0\right) \\ &\leq e^{-cnD^2(p_0||p_1)} . \end{aligned}$$

Thus the probability of false positive error is bounded in terms of 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 a false negative error can be found in a similar fashion:

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