Information Theory with Applications, Math6397
Lecture Notes from September 4, 2014

taken by Robert P Mendez

1.3 Neyman-Pearson, continued

1.3.5 Remark. We have seen that choosing an acceptance region based on a threshold value for the the likelihood ratio is optimal for the Neyman Pearson test problem. The asymmetry between the treatment of errors of the first and second kind will show up in the performance of this test, which is governed by the relative entropy, that is also not symmetric in both probability measures. To see this, we begin by reframing divergence as expected value to make use of the properties of that function.

1.3.1 Divergence Measures Test Effectiveness

Recall that the expected value of a random variable can be seen as a weighted average of outcomes. More specifically, if X is a random variable with alphabet $\mathbb{A}$ and associated probability $\mathbb{P}_X$, the expected value of X is $\mathbb{E}_X[X] = \sum_{a\in\mathbb{A}} \mathbb{P}_X(a)a$. If we consider $f(a) := \ln \frac{\mathbb{P}_X(a)}{\mathbb{P}_Y(a)}$ (where $\mathbb{P}_Y$ is as usual), then we may say $\mathbb{E}_X[f \circ X] = \sum_{a\in\mathbb{A}} \mathbb{P}_X(a)f(a)$ is the expected value of $f(X)$ and we have

$$D(X_1, X_2, \ldots, X_n \| \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n) = \mathbb{E}_{X_1,X_2,\ldots,X_n}\left[\ln \frac{\mathbb{P}_{X_1,X_2,\ldots,X_n}(X_1, X_2, \ldots, X_n)}{\mathbb{P}_{\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_n}(X_1, X_2, \ldots, X_n)}\right]$$

that is, divergence measures the expected value of the log of the ratio of the respective probabilities. We note that, since the X_i's and $\hat{X}_i$'s are independent, for any $x \in \mathbb{A}^n$,

$$\ln \frac{\mathbb{P}_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}{\mathbb{P}_{\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_n}(x_1, x_2, \ldots, x_n)} = \ln \frac{\mathbb{P}_{X_1}(x_1) \cdot \cdots \cdot \mathbb{P}_{X_n}(x_n)}{\mathbb{P}_{\hat{X}_1}(\hat{x}_1) \cdot \cdots \cdot \mathbb{P}_{\hat{X}_n}(x_n)} .$$

It follows by the product-to-sum property of the log and the linearity of the expected value function that

$$\begin{aligned}\mathbb{E}_{X_1,X_2,\ldots,X_n}\left[\ln \frac{\mathbb{P}_{X_1,X_2,\ldots,X_n}(X_1, X_2, \ldots, X_n)}{\mathbb{P}_{\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_n}(X_1, X_2, \ldots, X_n)}\right] &= \mathbb{E}_{X_1,X_2,\ldots,X_n}\left[\sum_{j=1}^{n} \ln \frac{\mathbb{P}_{X_j}(X_j)}{\mathbb{P}_{\hat{X}_j}(X_j)}\right] \\ &= \sum_{j=1}^{n} \mathbb{E}_{X_j}\left[\ln \frac{\mathbb{P}_{X_j}(X_j)}{\mathbb{P}_{\hat{X}_j}(X_j)}\right]\end{aligned}$$

which, since all the X_j's and $\hat{X}_j$'s were from the same distributions (and so identically distributed, and $\mathbb{P}_{X_i}(x) = \mathbb{P}_{X_j}(x)$ for all $x \in \mathbb{A}$, i and j), equals

$$\begin{aligned} &= \sum_{j=1}^{n} \mathbb{E}_{X_1}\left[\ln \frac{\mathbb{P}_{X_1}(x_1)}{\mathbb{P}_{\hat{X}_1}(x_1)}\right] \\ &= nD(X_1 \| \hat{X}_1) \end{aligned}$$

Recall, if $Y_1, Y_2, \ldots, Y_n$ are i.i.d., then $T_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ concentrates near its average. We have then, by the weak law of large numbers, that for $\epsilon > 0$

$$\mathbb{P}\left(|T_n - \mathbb{E}[T_n]| > \epsilon\right) \xrightarrow{n\to\infty} 0\,.$$

In our case, the summands Y_j are $\ln \frac{\mathbb{P}_{X_j}(X_j)}{\mathbb{P}_{\hat{X}_j}(X_j)}$, and as $n \to \infty$, $T_n = \frac{1}{n} \ln \frac{\mathbb{P}_{X_1,X_2,\ldots,X_n}}{\mathbb{P}_{\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_n}}$ is just an average and concentrates over $\mathbb{E}_{X_1,X_2,\ldots,X_n}[T_n]$, which equals $D(X_1\|\hat{X}_1)$. This limit of T_n is a good comparison for our choice of τ.

To have a meaningful test, τ must be chosen so that $\frac{1}{n}\ln\tau$ is close to the value where T_n concentrates, the divergence $D(X\|\hat{X})$ between the two random variables. We see, from the definition of τ, that a large value indicates good test performance, while a small value means poor performance. By our concentration argument, *divergence measures how successful the test is*–we need a large $D(X\|\hat{X})$ to distinguish the two measures reliably.

1.3.2 Relative Entropy and Data Processing

Another aspect of relative entropy that crops up in application involves questions we may have in data processing. We may "throw away information" by partitioning outcomes–grouping them, perhaps, and just keeping the information about the group to which each outcome belongs. What should we expect of the entropy of our product?

More formally, imagine an alphabet $\mathbb{A}$ is partitioned, each set receiving a label from a new alphabet $\{1, 2, \ldots, t\}$:

$$\mathbb{A} = \bigcup_{i=1}^{t} A_i,\ A_i \cap A_j = \emptyset\ \forall i \neq j$$

Our question can now be more clearly stated.

1.3.6 Question. How does the entropy of the induced measure on the new alphabet relate to the entropy of a random variable with alphabet $\mathbb{A}$? Does entropy increase, as it appears that we have "lost" information? On the other hand, if $t = 1$, entropy is now zero, as there is exactly one outcome.

1.3.7 Answer. In general, the two entropies cannot be compared. However, it turns out that the *relative* entropy increases, which we may see intuitively as resulting from the loss of information making the random variables "look more alike."

1.3.8 Lemma (Data Processing Inequality)**.** *Given X, Y with induced probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a common alphabet $\mathbb{A}$ and a partition $\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_t\}$ of $\mathbb{A}$, then*

$$D(\mathbb{P}\|\mathbb{Q}) = \sum_{a\in\mathbb{A}} \mathbb{P}(a) \ln \frac{\mathbb{P}(a)}{\mathbb{Q}(a)}$$

$$\geq \sum_{i=1}^{t} \mathbb{P}(\mathbb{A}_i) \ln \frac{\mathbb{P}(\mathbb{A}_i)}{\mathbb{Q}(\mathbb{A}_i)}$$

1.3.9 Remark. The reader may note that $\mathbb{P}(\mathbb{A}_i)$ is the probability of getting an a in the i^{th} subset.

Proof. Begin by converting divergence into a double sum:

$$\begin{aligned} D(X\|Y) &= \sum_{i=1}^{t} \sum_{a\in\mathbb{A}_i} \mathbb{P}(a) \ln \frac{\mathbb{P}(a)}{\mathbb{Q}(a)} \\ &\overset{log\text{–}sum}{\geq} \sum_{i=1}^{t} \left(\sum_{a\in\mathbb{A}} \mathbb{P}(a) \right) \ln \frac{\sum\limits_{a\in\mathbb{A}_i} \mathbb{P}(a)}{\sum\limits_{a\in\mathbb{A}_i} \mathbb{Q}(a)} \end{aligned}$$

Noting that $\sum\limits_{a\in\mathbb{A}_i} \mathbb{P}(a)$ is exactly the probability $\mathbb{P}(\mathbb{A}_i)$, this gives

$$D(X\|Y) \geq \sum_{i=1}^{t} \mathbb{P}(\mathbb{A}_i) \ln \frac{\mathbb{P}(\mathbb{A}_i)}{\mathbb{Q}(\mathbb{A}_i)}\,.$$

□

1.3.10 Question. Prompted by analysis instincts, consider the following: If $D(\mathbb{P}\|\mathbb{Q})$ is small, how "close" are $\mathbb{P}$ and $\mathbb{Q}$?

1.3.11 Answer. The following proposition ties the sum of the distances between $\mathbb{P}(a)$ and $\mathbb{Q}(a)$ over the entire alphabet.

1.3.12 Proposition (Pinsker)**.** *For discrete random variables X and Y with common alphabet $\mathbb{A}$,*

$$\begin{aligned} D(\mathbb{P}\|\mathbb{Q}) &\geq \frac{1}{2}\|\mathbb{P}-\mathbb{Q}\|_1^2 \\ &= \frac{1}{2}\left(\sum_{a\in\mathbb{A}} |\mathbb{P}(a)-\mathbb{Q}(a)| \right)^2 \end{aligned}$$

where $\|\cdot\|_1$ is the L^1 norm

1.3.13 Remark. Our method will be to reduce the alphabet and then show that it is enough to prove the reduced case.

Proof. Partition $\mathbb{A}$ by defining $\mathbb{A}_0 := \{a \in \mathbb{A} \mid \mathbb{P}(a) \geq \mathbb{Q}(a)\}$ and $\mathbb{A}_1 := \mathbb{A} \setminus \mathbb{A}_0$. Then, by the data processing inequality, we know

$$D(\mathbb{P}\|\mathbb{Q}) \geq \sum_{i=0}^{1} \mathbb{P}(\mathbb{A}_i) \ln \frac{\mathbb{P}(\mathbb{A}_i)}{\mathbb{Q}(\mathbb{A}_i)}$$

If we let $\hat{\mathbb{P}}(i) := \mathbb{P}(\mathbb{A}_i)$ and $\hat{\mathbb{Q}}(i) := \mathbb{Q}(\mathbb{A}_i)$, then

$$D(\mathbb{P}\|\mathbb{Q}) \geq D(\hat{\mathbb{P}}\|\hat{\mathbb{Q}})$$

On the other hand,

$$\begin{aligned}
\|\mathbb{P}-\mathbb{Q}\|_1 &= \sum_{a\in\mathbb{A}} |\mathbb{P}(a)-\mathbb{Q}(a)| \\
&= \sum_{a\in\mathbb{A}_0} (\mathbb{P}(a)-\mathbb{Q}(a)) - \sum_{a\in\mathbb{A}_1} (\mathbb{P}(a)-\mathbb{Q}(a)) \\
&= |\mathbb{P}(\mathbb{A}_0)-\mathbb{Q}(\mathbb{A}_0)| + |\mathbb{P}(\mathbb{A}_1)-\mathbb{Q}(\mathbb{A}_1)| \\
&= \|\hat{\mathbb{P}}-\hat{\mathbb{Q}}\|_1
\end{aligned}$$

$a \in \mathbb{A}_i$ determines sign of $\mathbb{P}(a)-\mathbb{Q}(a)$

Since $\|\mathbb{P}-\mathbb{Q}\|_1 = \|\hat{\mathbb{P}}-\hat{\mathbb{Q}}\|_1$, we see that it is enough to prove the inequality for the binary case, as our preparatory remark foretold. Abbreviating, let $p := \mathbb{P}(\mathbb{A}_0)$ and $q := \mathbb{Q}(\mathbb{A}_0)$, and we have that

$$D(\hat{\mathbb{P}}\|\hat{\mathbb{Q}}) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$$

and

$$\begin{aligned}
\|\hat{\mathbb{P}}-\hat{\mathbb{Q}}\|_1^2 &= \left(|p-q| + |(1-p)-(1-q)|\right)^2 \\
&= 4(p-q)^2
\end{aligned}$$

By making appropriate substitutions and rearranging $D(\mathbb{P}\|\mathbb{Q}) \geq \frac{1}{2}\|\mathbb{P}-\mathbb{Q}\|_1^2$, we now need only to prove that

$$f(p,q) := p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q} - 2(p-q)^2 \geq 0$$

First, we fix p and look for extrema. Note that p and q take on values between 0 to 1; endpoint tests reveal that as $q \to 1$ or $q \to 0$, $f \to \infty$. Looking for critical points, set

$$\frac{\partial}{\partial q} f(p,q) = -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) = 0$$

which reduces to a factored form of

$$\left(4 - \frac{1}{q(1-q)}\right)(p-q) = 0$$

And we have critical points $q = \frac{1}{2}$ and $q = p$. If $q = p$, $f(p,p) = 0$, which is nonnegative, as desired. If $q = \frac{1}{2}$, we consider$f(p, \frac{1}{2})$ and, again, search for extrema. Our endpoint test reveals positive values at $p = 0$ and $p = 1$. Critical points:

$$\frac{\partial}{\partial p} f(p, \frac{1}{2}) = \ln(p) + 1 - \ln(1-p)1 - 4(p - \frac{1}{2}) = 0$$

reduces to $\frac{p}{1-p} = e^{-4p+2}$. The left-hand side is increasing on $(0,1)$, while the right-hand side is decreasing, so there is at most one solution. It turns out that $p = \frac{1}{2}$ solves it, which means $p = q$

and $f(p,q) = 0$ is non-negative. Since all critical points give non-negative values, we have that $f(p,q) \geq 0$.

□