Bulletin of the American Mathematical Society (Similar)

The Author

The Date

Abstract. Replace this text with your own abstract.

1. The Neyman-Pearson Region

Given two distributions P_0 and P_1, (and methods for sampling from each), we can randomly sample a measurable set S in the sample space and approximate the two probabilities

$$\alpha = P(x \in S | x \sim P_1) \tag{1}$$

$$\beta = P(x \notin S | x \sim P_0). \tag{2}$$

As inttroduced by Neyman and Pearson (1928) these are the probabilities of misclassifying x according to it's origin. According to the famous lemma of Neyman and Pearson (1933), the region R of the unit square composed of pairs (α, β) which arise in this way is convex (because we can randomize tests), includes the 'line of ignorance' $\alpha + \beta = 1$, and is symmetric about the point $\left(\frac{1}{2}, \frac{1}{2}\right)$ (because we can complement tests).

Remark 1. *The idea that we will be considering only two possible explanations for an observation is an extreme simplification, although it is valuable in classification problems, Many examples of hypothesis testing involve 'complex hypotheses' which are beyond the scope of this note..*

The boundary of R has two branches; the lower branch of the boundary are the tests which have the least β for any given α, and the reflection of this boundary in $\left(\frac{1}{2}, \frac{1}{2}\right)$, which correspond to the highest β for any give α. For the distributions we consider, the most powerful tests will be given by the Neyman-Pearson lemma, which determines the lower boundary of R.

The convex hull of the pairs (α, β) obtained by random sampling sets S is a subset of R; how the rate at which R is exhausted in this way depends largely on the dimension of the sample space; higher dimensional space requiring many more sets to sample.

If P_0 and P_1 belong to an exponential family, then the log likelihood ratio is a linear function of the sufficient statistics; the exponential family transforms classification from the 'feature set' of the sample space to classification using the sufficient statistics as features. The property of sufficiency is equivalent to the property that classification in the sample space can be exactly solved in the coordinates of the sufficient statistics. The Darmois Koopman Pitman theorem implies that outside of exponential families, finite dimensional classification is only approximate.

The Neyman-Pearson family of tests are indexed by a parameter λ, we consider $\alpha(\lambda)$ and $\beta(\lambda)$, so we can approach the total variation distance calculation as the minimization

$$s = \min \alpha(\lambda) + \beta(\lambda) \tag{3}$$

which in view of the convexity of R implies that this is a convex minimization. We know that $(0, 1)$ and $(1, 0)$ are points in R. The total variation distance is related to s by

$$1 - \frac{1}{2} TVD = s \tag{4}$$

Example 2. *In the concentric (e.g. zero mean) Gaussian case the calculation of the Neyman-Pearson tests is tractable, because this is a family with a low dimensional sufficient statistic. In the concentric Gaussian case, the region R is a function the eigenvalues of the covariance ratio $C_1^{-1}C_2^2C_1^1$, and is also invariant under inversion of any of these eigenvalues; and any function of R such as TVD, (or s which is the Euclidean distance from R to 0) is determined as a function of these eigenvalues.*

Remark 3 ['Receiver Operating Characteristic']**.** *The Neyman-Pearson region is closely related to the idea of the receiver operating characteristic (ROC) since $\alpha(S)$ is the false positive rate of the test S and $\beta(S)$ is the false negative rate, hence $1-\beta(S)$ is the 'true positive rate'. So a receiver operating 'point' is a point $(\alpha(S), 1-\beta(S))$. For a one parameter family of sets S_t, then the receiver operating characteristic is the curve $(\alpha(S_t), 1-\beta(S_t))$. If we reflect the Neyman-Pearson region in the horizontal line $\beta = 1/2$, then the Neyman-Pearson boundary is an upper bound for any receiver operating characteristic.*

1.1. Area A and Perimeter L. The total variation distance an be defined as the minimum distance between two lines of slope -1 which line on opposite sides of R.

Since R is convex, it contains a triangle with vertices $(0,1)$, $(1,0)$ and (ξ,η), where $\xi+\eta=s$. Such a triangle has base $\sqrt{2}$ and height $\frac{1}{\sqrt{2}}(1-s)$, and area $\frac{1}{2}(1-s)$. R also contains the reflection of this triangle. R is a subset of the hexagon with area $1-s^2$ which results from removing the triangle $\{0 \le x \le s,\, 0 \le y \le s,\, x+y \le s\}$ and its reflection in $(1/2, 1/2)$ from the unit square. (The triangles each have area $\frac{1}{2}s^2$ and are disjoint).

These imply

$$1-s \le A \le 1-s^2 \tag{5}$$

and using

$$\frac{1}{2}TVD = 1-s \tag{6}$$

and

$$1+s = 2-(1-s) \tag{7}$$

we obtain

$$\frac{1}{2}TVD \le A \le TVD - \frac{1}{4}(TVD)^2 \tag{8}$$

so that the area is within a factor of 2 of the TVD.

Remark 4. *The 'ROC AUC', that is, the area under a receiver operating characteristic depends on the family of tests S_t, but the upper bound on this area for 'all' families is $\frac{1}{2}(1+A)$ which is attained for the Neyman-Pearson tests. The bound on the Neyman-Pearson area translates to*

$$\frac{1}{2}\left(1+\frac{1}{2}TVD\right) \le \frac{1}{2}(1+A) \le \frac{1}{2}\left(1+TVD-\frac{1}{4}(TVD)^2\right) \tag{9}$$

$$\frac{1}{2}+\frac{1}{4}TVD \le \frac{1}{2}(1+A) \le \frac{1}{2}+\frac{1}{2}TVD-\frac{1}{8}(TVD)^2 \tag{10}$$

note that the TVD does NOT in general provide any lower bound on the 'ROC AUC' for some test, because a test may be suboptimal. However 'ROC AUC' does provide a lower bound for TVD.

Exercise 5. *Show that* $A(p,q) = A(q,p)$.

Exercise 6. *Use the inclusion relations between the hexagon and triangles mentioned above to obtain upper and lower bounds on the perimeter* L *of the Neyman-Pearson region.*

Exercise 7. *Let* p *be the density for the normal distribution on the line with mean* 0 *and variance* 1*, and* q *the density for the normal distribution on the line with mean* 0 *and variance* 9*. Find* $TVD(p,q)$ *analytically. Show that the Neyman Pearson region for* p *and* q *is the same as for* p' *and* q'*, where* p' *is the density of the normal distribution in* d *dimensions with covariance* I *and* q' *is the density of the normal distribution with covariance* D*, where* D *is diagonal and* $D_{11} = 9$ *and* $D_{jj} = 1$ *for* $j > 1$.

2. Divergence Approach

We have two densities p and q with respect to a measure μ and for each measurable set E we define

$$\alpha(E) = \int_E p d\mu \tag{11}$$

$$\beta(E) = 1 - \int_E q d\mu. \tag{12}$$

For each nonnegative u, v we write $\rho = \min_E u\alpha + v\beta$.

Exercise 8. *Show that the line in the* (α, β) *plane* $u\alpha + v\beta = \rho$ *is tangent to the Neyman-Pearson region* R *and the half plane* $u\alpha + v\beta \geq \rho$ *contains* R*. This line is called a supporting line for the convex set* R*.*

Observe that

$$\rho(u,v) = \min_E u \int_E p d\mu + v - v \int_E q d\mu \tag{13}$$

$$= v - \min_E \int_E (vq - up)\, d\mu \tag{14}$$

which is minimized by choosing E to be the level set $\{x \text{ s.t. } vq(x) - up(x) \leq 0\}$. This is the same as the level set of the 'log likelihood ratio'

$$E = \left\{ x \text{ s.t. } \log \frac{p(x)}{q(x)} \geq \log \frac{v}{u} \right\} \tag{15}$$

which corresponds to a Neymann-Pearson test. If $p = q$ (a.e. x) then $\alpha + \beta = 1$.

We have

$$\rho(u,v) = v - \int (vq - up)_+ \, d\mu \tag{16}$$

where $(x)_+$ is the positive part of x, equivalently $(x)_+ = \frac{1}{2}(x + |x|)$.

Exercise 9. *Show that*

$$\frac{1}{2} TVD = 1 - \min \alpha(\lambda) + \beta(\lambda) = \frac{1}{2} \int_E |q - p| \, d\mu. \tag{17}$$

Definition 10. *An f-divergence $D_f(p||q)$ is an expectation*

$$D_f(p||q) = \int qf\left(\frac{p}{q}\right) d\mu \tag{18}$$

where f is convex and $f(1) = 0$.

Exercise 11. *What does Jensen's inequality say about $D_f(p||q)$?*

Exercise 12. *Show that a convex combination $tD_{f_1}(p||q) + (1-t) D_{f_2}(p||q)$ where $0 \leq t \leq 1$ defines an f-divergence.*

Exercise 13. *The* perspective *g of a function f is defined as $g(x,y) = yf\left(\frac{x}{y}\right)$. For this exercise assume f and g are defined on nonnegative arguments. Show that g is convex if and only if f is. (Hint: $f(u) = g(1,u)$.) Show that $f(1) = 0$ is equivalent to $g(x,x) = 0$. Conclude that every f-divergence can be defined as*

$$D_f(p||q) = \int g(p,q)\, d\mu \tag{19}$$

where g is convex and $g(p,p) = 0$ for all p. Show that the 'reverse' f-divergence $D_f(q||p)$ must also be an f-divergence, as does any convex combination

$$tD_f(p||q) + (1-t) D_f(p||q). \tag{20}$$

3. The Second Computation Project

1. Consider the integral
$$\rho(u,v) = v - \int (vq - up)_+ \, d\mu \tag{21}$$
from the point of view of Monte Carlo; if samples are drawn from p then we have
$$\rho(u,v) = v\left(1 - \int p\left(\frac{q}{p} - \frac{u}{v}\right)_+ d\mu\right) \tag{22}$$
and if they are drawn from q we have
$$\rho(u,v) = u\left(1 - \int q\left(\frac{v}{u} - \frac{p}{q}\right)_+\right) d\mu. \tag{23}$$
Give an expression which represents samples drawn from a mixture $tp + (1-t)q$.

2. Let p be the density for the normal distribution on the line with mean 0 and variance 1, and q the density for the normal distribution on the line with mean 0 and variance 9. Find $TVD(p,q)$ analytically. Show that the Neyman Pearson region for p and q is the same as for p' and q', where p' is the density of the normal distribution in d dimensions with covariance I and q' is the density of the normal distribution with covariance D, where D is diagonal and $D_{11} = 9$ and $D_{jj} = 1$ for $j > 1$. Estimate and plot the supporting lines of the Neyman-Pearson region by Monte Carlo for a variety of values (u,v). Include

3. Repeat the previous exercise for the concentric ellipsoidal Laplace and Cauchy distributions. Compare the Neyman-Pearson regions to the Gaussian case.

4. Discuss the sample size and error for your estimates.

5. Can you find a variance reduction method?