Discriminative and Generative Views of Binary Experiments

Mark D. Reid
Australian National University
Canberra ACT 0200, Australia
Mark.Reid@anu.edu.au

Robert C. Williamson
Australian National University and NICTA
Canberra ACT 0200, Australia
Bob.Williamson@anu.edu.au

Abstract

We consider Binary experiments (supervised learning problems where there are two different labels) and explore formal relationships between two views of them, which we call "generative" and "discriminative". The discriminative perspective involves an expected loss. The generative perspective (in our sense) involves the distances between class-conditional distributions. We extend known results to the class of all proper losses (scoring rules) and all f-divergences as distances between distributions. We also sketch how one can derive the SVM and MMD algorithms from the generative perspective.

This brief note summarises some recent results on two views of binary experiments. Details, proofs and references to the literature are available in a preprint[1] which we refer to.

A binary experiment comprises two class conditional distributions P and Q (corresponding to positive and negative labelled examples) and a prior probability $\pi \in (0,1)$ which is the probability of a positive example being drawn. We have studied different views of such experiments: optimising a expected risk defined in terms of a proper loss, and measuring the distance between P and Q using f-divergences. A classical result (for $\pi = \frac{1}{2}$) if that

$$\underline{\mathbb{L}}_{0-1}(P,Q) = \frac{1}{2} - \frac{1}{4} V(P,Q);$$

that is the minimal (i.e. Bayes) 0-1 risk for an experiment (P,Q) is $\frac{1}{2}$ minus a quarter of the variational divergence between the distributions. This makes clear intuitive sense — the further apart P and Q are, the easier the discrimination problem. The results we derive are more general than existing results. We pay particular attention to the case of general $\pi \in (0,1)$ rather than the common special case of $\pi = \frac{1}{2}$. We present analogs of the above result for arbitrary proper losses and arbitrary f-divergences.

An alternative parametrisation of a binary experiment is (η, M) where M is the (unconditional) distribution of examples over the input space and η is the conditional distribution $\eta(x) = \Pr(Y = 1|X = x)$. There is a simple transformation between (π, P, Q) and (η, M).

A key result of the preprint is Theorem 10.

[1] Mark D. Reid and Robert C. Williamson, "Information, Divergence and Risk for Binary Experiments," August 2009 http://axiom.anu.edu.au/∼williams/InformationDivergenceRisk.pdf

Theorem 1 *Let $f : [0, \infty) \to \mathbb{R}$ be a convex function and for each $\pi \in [0, 1]$ define for $c \in [0, 1)$:*

$$\begin{aligned}\phi(c) &:= \frac{1-c}{1-\pi} f\left(\lambda_\pi(c)\right) && (1)\\ \underline{L}(c) &:= -\phi(c) && (2)\end{aligned}$$

where λ_π is particular function (defined in the paper). Then for every binary experiment (P, Q) we have

$$\mathbb{I}_f(P, Q) = \Delta\underline{\mathbb{L}}(\eta, M) = \mathbb{B}_\phi(\eta, M) \tag{3}$$

where $M := \pi P + (1 - \pi)Q$, $\eta := \pi dP/dM$ and $\underline{\mathbb{L}}$ is the expectation (in X) of the conditional Bayes risk $\underline{L}$.

This theorem says that given a binary experiment with class conditional distributions P and Q, one can define convex functions ϕ in terms of a chosen f which means that the f divergence $\mathbb{I}_f(P, Q)$ between P and Q equals the statistical information $\Delta\underline{\mathbb{L}}(\eta, M)$ which equals the generative Bregman divergence.

The statistical information

$$\Delta\underline{\mathbb{L}}(\eta, M) = \Delta\underline{\mathbb{L}}(\pi, P, Q) := \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M)$$

where the right side is the difference between the prior and posterior Bayes risk. (The notion is due to Degroot). The generative Bregman divergence is

$$\mathbb{B}_\phi(P, Q) := \mathbb{E}_M\left[B_\phi(p, q)\right] = \mathbb{E}_{\mathsf{X}\sim M}\left[B_\phi(p(\mathsf{X}), q(\mathsf{X}))\right].$$

where B_ϕ is a standard Bregman divergence with respect to the convex function ϕ.

Elsewhere in the preprint (section 5) we show how f-divergences and Statistical Information can be represented in terms of integral representations — weighted one dimensional integrals of primitives. For Statistical Information the primitives are cost sensitive misclassification losses. For f-divergences they are particular f-divergences related to the Variational divergence. We show how the weight function for f-divergence that corresponds to a loss through the relationship with statistical information in the above theorem can be explicitly given as a function of the weight function for the representation of the loss.

The final result that may be of interest to the workshop is in Appendix F of the preprint. There we make use of the relationship between a variant of variational divergence and 0-1 loss. We show how one can derive the support vector machine from the "generative" perspective. One starts with the two class conditional distributions P and Q. One estimates a variant of the variational divergence between P and Q. Then by a natural (but novel) inductive principle one can derive the SVM in a new and simple manner. In doing so one sees that the recently popular Maxmimum Mean Discrepancy (a method of measuring distance between distributions) simply corresponds to a solution of the classification problem via the kernel Fisher discriminant.

Thus we show the value of looking at a problem in two different ways: one can either say one has two distributions P and Q and one is interested in how close they are, or one can interpret P and Q as class conditional distributions and then solve a classification problem. These are two sides of the same coin.