Discriminative and Generative Views of Binary Experiments

Robert Williamson

December 2009

THE AUSTRALIAN NATIONAL UNIVERSITY

NICTA

Joint work with Mark Reid

Statistical Experiments with Two Distributions

Consider the classical result

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

where

$$\underline{\mathbb{L}}^{0-1}(\tfrac{1}{2}, P, Q) = \inf_{r \in \{0,1\}^{\mathcal{X}}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathbb{P}}[\ell^{0-1}(r(\mathsf{X}), \mathsf{Y})].$$

is the Bayes risk with respect to 0-1 loss for a classification problem with class conditional distributions P and Q and *a priori* probability of a positive label $\frac{1}{2}$ and

$$V(P, Q) = 2 \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)| = \int_{\mathcal{X}} |p(x) - q(x)| dx$$

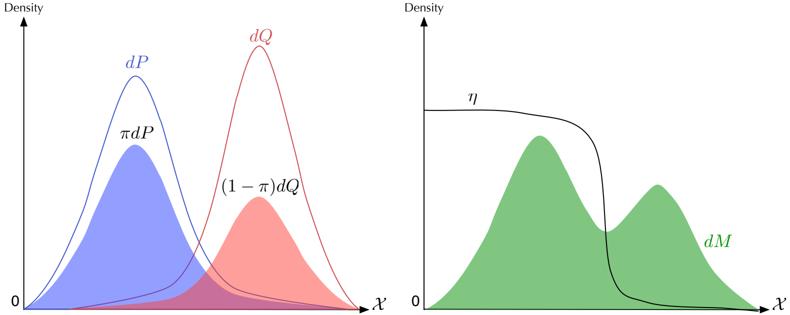
is the Variational Divergence between distributions P and Q.

Theme of this talk

Generalisations and implications of (1).

Details in paper — see workshop or my webpage.

Generative and Discriminative Perspectives



Translating between the perspectives

$$M = \pi P + (1-\pi)Q \text{ and } \eta = \pi \frac{dP}{dM}$$

Loss Functions

- A function $\hat{\eta} : \mathcal{X} \to [0,1]$ is a class probability *estimator*.
- Also $\hat{\eta} = \hat{\eta}(x) \in [0,1]$ denotes an *estimate* for a specific observation.
- Estimate quality is assessed using a *loss function*

$$\ell : \{0,1\} \times [0,1] \to \mathbb{R}$$

- The loss of the estimate $\hat{\eta}$ with respect to the label $y \in \mathcal{Y}$ is denoted $\ell(y, \hat{\eta})$.
- If $\eta \in [0,1]$ is the probability of observing the label $y = 1$ the cost-weighted *point-wise risk* of the estimate $\hat{\eta} \in [0,1]$ is defined to be the η-average of the point-wise loss for $\hat{\eta}$:

$$L(\eta, \hat{\eta}) := \mathbb{E}_{\mathsf{Y} \sim \eta}[\ell(\mathsf{Y}, \hat{\eta})] = \ell(0, \hat{\eta})(1 - \eta) + \ell(1, \hat{\eta})\eta.$$

Loss Functions (continued)

- When $\eta : \mathcal{X} \to [0,1]$ is an observation-conditional density, taking the M-average of the point-wise risk gives the *(full) risk*

$$\mathbb{L}(\eta, \hat{\eta}, M) := \mathbb{E}_{\mathsf{X} \sim M}[L(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))] = \int_{\mathcal{X}} L(\eta(x), \hat{\eta}(x))\, dM(x)$$

- ℓ, L and $\mathbb{L}$ denote loss, point-wise and full risk of $\hat{\eta}$:
- The combination of a loss ℓ and the distribution $\mathbb{P}$ is a *task*.
- Discriminatively $T = (\eta, M; \ell)$; Generatively $T = (\pi, P, Q; \ell)$.
- A natural measure of the difficulty of a task is its minimal achievable risk, or *Bayes risk*:

$$\underline{\mathbb{L}}(\eta, M) = \underline{\mathbb{L}}(\pi, P, Q) := \inf_{\hat{\eta}: \mathcal{X} \to [0,1]} \mathbb{L}(\eta, \hat{\eta}, M) = \mathbb{E}_M\left[\underline{L}(\eta)\right],$$

where

$$[0,1] \ni \eta \mapsto \underline{L}(\eta) := \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta})$$

is the *point-wise Bayes risk*.

Generalising 0-1 loss: Proper Losses

- *Proper losses* are losses for probability estimation that have a point-wise risk $L(\eta, \hat{\eta})$ that is minimised when $\hat{\eta} = \eta$.
- A proper loss ℓ satisfies $\underline{L}(\eta) = L(\eta, \eta)$ for all $\eta \in [0, 1]$.

We consider all proper losses.

Savage's Theorem

A proper loss can be expressed in terms of its conditional Bayes risk:

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Generalising Variational Divergence: f-divergences

Definition

The *f-divergence of P from Q* is

$$\mathbb{I}_f(P,Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

f-Divergence Examples

Symbol	Divergence Name	$f(t)$
V	Variational	$\lvert t-1 \rvert$
KL	Kullback-Liebler	$t \ln t$
Δ	Triangular Discrimination	$(t-1)^2/(t+1)$
I	Jensen-Shannon	$\frac{t}{2} \ln t - \frac{(t+1)}{2} \ln(t+1) + \ln 2$
T	Arithmetic-Geometric Mean	$\left(\frac{t+1}{2}\right) \ln \left(\frac{t+1}{2\sqrt{t}}\right)$
J	Jeffreys	$(t-1)\ln(t)$
h^2	Hellinger	$(\sqrt{t}-1)^2$
χ^2	Pearson χ-squared	$(t-1)^2$
Ψ	Symmetric χ-squared	$\frac{(t-1)^2(t+1)}{t}$

A Key Result

Theorem

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

$$\phi(c) := \frac{1-c}{1-\pi} f\left(\lambda_\pi(c)\right), \qquad \underline{L}(c) := -\phi(c)$$

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

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

where $M := \pi P + (1-\pi) Q$ *and* $\eta := \pi dP/dM$*.*

Given a binary experiment with class conditional distributions P and Q, one can define convex functions ϕ in terms of a chosen f such 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 $\mathbb{B}_\phi(\eta, M)$.

Where. . .

The statistical information is the difference between the prior and posterior Bayes risk:

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

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.

Integral Representations

- All proper losses can be written as a weighted integral of primitive losses (cost-sensitive misclassification losses)
- All f-divergences can be written as a weighted integral of primitive f-divergences (generalisations of the variational divergence)
- The corresponding weight functions are a much nicer parametrisation
- There is a direct correspondence between the respective weight functions (as a corollary of the previous theorem)
- The integral representations are useful for other things
  - Surrogate regret bounds [ICML2009]
  - Generalised Pinsker Inequalities [COLT2009]

A Curious Development from the New Perspective

- Consider the following generalisation of V:

$$V_{\mathcal{R},\pi}(P,Q) := 2 \sup_{r \in \mathcal{R} \subseteq [-1,1]^{\mathcal{X}}} |\pi \mathbb{E}_P r - (1-\pi)\mathbb{E}_Q r|,$$

  where $\pi \in (0,1)$.
- Consider the linear loss

$$\ell^{\text{lin}}(r(x),y) := 1 - yr(x), \quad y \in \{-1,1\}.$$

- If r is unrestricted, then there is no guarantee that $\ell^{\text{lin}} > -\infty$ and is thus a legitimate loss function.
- Below we will always consider $r \in \mathcal{R}$ such that the linear loss is bounded from below.

Relationship between $V_{\mathcal{R},\pi}(P,Q)$ and $\underline{\mathbb{L}}^{\mathrm{lin}}_{\mathcal{R}}(\pi,P,Q)$

Theorem

Assume that $\mathcal{R} \subseteq [-a,a]^{\mathcal{X}}$ for some $a > 0$ and is symmetric about zero. Then for all $\pi \in (0,1)$ and all distributions P and Q on $\mathcal{X}$

$$\underline{\mathbb{L}}^{\mathrm{lin}}_{\mathcal{R}}(\pi,P,Q) = 1 - \frac{1}{2} V_{\mathcal{R},\pi}(P,Q)$$

and the r that attains $\underline{\mathbb{L}}^{\mathrm{lin}}_{\mathcal{R}}(\pi,P,Q)$ corresponds to the r that obtains the supremum in the definition of $V_{\mathcal{R},\pi}(P,Q)$.

In an RKHS

- Suppose that $\mathcal{R} = B_{\mathcal{H}} := \{r : \|r\|_{\mathcal{H}} \leq 1\}$, the unit ball in $\mathcal{H}$, a Reproducing Kernel Hilbert Space.
- Thus for all $r \in \mathcal{R}$ there exists a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ such that $r(x) = \langle r, \phi(x) \rangle_{\mathcal{H}}$ and $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y)$, where k is a positive definite *kernel* function.
- Borgwardt et al. show that

$$V^2_{B_{\mathcal{H}}, \frac{1}{2}}(P, Q) = \frac{1}{4}\|\mathbb{E}_P \phi - \mathbb{E}_Q \phi\|^2_{\mathcal{H}}.$$

- Thus

$$\mathbb{L}^{\mathrm{lin}}_{\mathcal{R}}(\pi, P, Q) = 1 - \frac{1}{4}\|\mathbb{E}_P \phi - \mathbb{E}_Q \phi\|_{\mathcal{H}}.$$

Empirical Estimators of $V_{\mathcal{R},\pi}(P,Q)$

- Given an independent identically distributed sample $\mathbf{w} = (w_1, \ldots, w_m) \in \mathcal{X}^m$ the $\boldsymbol{\alpha}$-weighted empirical distribution $\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}}$ with respect to $\mathbf{w}$ is defined by

$$d\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}} := \sum_{i=1}^{m} \alpha_i \delta(\cdot - w_i)$$

  where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, $\alpha_i \geq 0$, $i = 1, \ldots, m$ and $\sum_{i=1}^{m} \alpha_i = 1$.
- We will write $\hat{\mathbb{E}}_{\mathbf{w}}^{\boldsymbol{\alpha}} \phi := \mathbb{E}_{\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}}} \phi = \sum_{i=1}^{m} \alpha_i \phi(w_i)$.
- Thus

$$V^2_{\mathcal{R},\frac{1}{2}}(\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}}, \hat{P}_{\mathbf{z}}^{\boldsymbol{\beta}}) = \frac{1}{2} \|\hat{\mathbb{E}}_{\mathbf{w}}^{\boldsymbol{\alpha}} \phi - \hat{\mathbb{E}}_{\mathbf{z}}^{\boldsymbol{\beta}}\|^2_{\mathcal{H}}.$$

Empirical Estimators

- P and Q correspond to the positive and negative class conditional distributions.
- Let $\mathbf{x} := (x_1, \ldots, x_m)$ be a sample drawn from $M = \pi P + (1 - \pi) Q$ with corresponding label vector $\mathbf{y} = (y_1, \ldots, y_m)$.
- $I := \{1, \ldots, m\}$, $I^+ := \{i \in I : y_i = 1\}$, $I^- := \{i \in I : y_i = -1\}$.
- Consider a weight vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$.
- Thus

$$\hat{\mathbb{E}}_P \phi = \sum_{i \in I^+} \alpha_i \phi(x_i) \quad \text{and} \quad \hat{\mathbb{E}}_Q \phi = \sum_{i \in I^-} \alpha_i \phi(x_i)$$

where $\sum_{i \in I^+} \alpha_i = \frac{m^+}{m}$ and $\sum_{i \in I^-} \alpha_i = \frac{m^-}{m}$ and hence $\sum_{i \in I} \alpha_i y_i = \frac{m^+ - m^-}{m}$.

Empirical Estimators (Cont.)

We have

$$\begin{aligned}
2V_{B_{\mathcal{H}},\frac{1}{2}}(\hat{P},\hat{Q}) &= \left\langle \hat{\mathbb{E}}_P\phi - \hat{\mathbb{E}}_Q\phi, \hat{\mathbb{E}}_P\phi - \hat{\mathbb{E}}_Q\phi \right\rangle \\
&= \left\langle \sum_{i\in I^+} \alpha_i\phi(x_i) - \sum_{i\in I^-} \alpha_i\phi(x_i), \sum_{j\in I^+} \alpha_j\phi(x_j) - \sum_{j\in I^-} \alpha_j \right\rangle \\
&= \left\langle \sum_{i\in I} \alpha_i y_i \phi(x_i), \sum_{j\in I} \alpha_j y_j \phi(x_j) \right\rangle \\
&= \sum_{i\in I}\sum_{j\in I} \alpha_i\alpha_j y_i y_j \langle \phi(x_i), \phi(x_j)\rangle \\
&= \sum_{i\in I}\sum_{j\in I} \alpha_i\alpha_j y_i y_j k(x_i, x_j) =: J(\boldsymbol{\alpha}, \mathbf{x}). \qquad (2)
\end{aligned}$$

We now consider three different choices of $\boldsymbol{\alpha}$.

Uniform Weighting

- If we set $\alpha_i = \frac{1}{m}$, $i = 1, \ldots, m$, then (2) becomes

$$\frac{1}{m^2} \sum_{i,j \in I} y_i y_j k(x_i, x_j) = \mathrm{MMD}_b^2[B_{\mathcal{H}}, \mathbf{x}^+, \mathbf{x}^-]$$

  where $\mathbf{x}^+ := (x_i)_{i \in I^+}$, $\mathbf{x}^- := (x_i)_{i \in I^-}$.
- MMD_b is the biased estimator of the *Maximum Mean Discrepancy*, an alternate name for $V_{\mathcal{R}}$.
- This case corresponds to using a Fisher linear discriminant in feature space when it is assumed that the within-class covariance matrices are both the identity matrix.

Pessimistic Weighting

- Instead of weighting each sample equally, one can optimise over $\boldsymbol{\alpha}$.
- Minimizing $J(\boldsymbol{\alpha}, \mathbf{x})$ over $\boldsymbol{\alpha}$ will maximize $\underline{\mathbb{L}}^{\mathrm{lin}}$ and is thus the most pessimistic choice:

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m}\sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) && (3)\\
\text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m && (4)\\
& \sum_{i=1}^{m} \alpha_i y_i = \frac{m^+ - m^-}{m} && (5)\\
& \sum_{i=1}^{m} \alpha_i = 1 && (6)
\end{aligned}$$

which can be recognized as the Support Vector Machine.

- The SVM uses the sign of the "witness" $x \mapsto \sum_{i=1}^{m} \alpha_i y_i k(x_i, x)$ as its predictor.

Interpolation Between Above Two Cases

- A parametrized interpolation between the above two cases can be constructed by the addition of the constraints

$$\alpha_i \leq \frac{1}{\nu m}, \quad i = 1, \ldots, m, \tag{7}$$

  where $\nu \in (0, 1]$ is an adjustable parameter.
- ν controls the sparsity of $\boldsymbol{\alpha}$ since (7), (4) and (6) together imply that $|\{i \in I : \alpha_i \neq 0\}| \geq \nu m$.
- Crisp and Burges have shown that (3),...,(7) is equivalent to the ν-SVM algorithm.

An Alternate Inductive Principle

- The traditional Empirical Risk Minimization principle replaces (P, Q) with $(\hat{P}_{\mathbf{x}^+}, \hat{Q}_{\mathbf{x}^-})$ in the definition of $\mathbb{L}(\pi, P, Q)$.
- Then, in order to not overfit, one restricts the class of functions from which hypotheses are drawn.

$$\mathbb{L}(\pi, P, Q) \xrightarrow{\text{Empirical Approximation (uniform)}} \mathbb{L}(\pi, \hat{P}_{\mathbf{x}^+}, \hat{Q}_{\mathbf{x}^-}) \xrightarrow{\text{Restrict Class}} \mathbb{L}_{\mathcal{R}}(\pi, \hat{P}_{\mathbf{x}^+}, \hat{Q}_{\mathbf{x}^-}).$$

- Set $\boldsymbol{\alpha}^+ = (\alpha_i)_{i \in I^+}$ and $\boldsymbol{\alpha}^- = (\alpha_i)_{i \in I^-}$.
- The derivation above corresponds to

$$\text{“}\mathbb{L}(\pi, P, Q)\text{”} \xrightarrow{\text{Restrict Class}} \mathbb{L}_{\mathcal{R}}(\pi, P, Q) \xrightarrow{\text{Empirical Approximation } (\boldsymbol{\alpha}\text{-weighted})} \mathbb{L}_{\mathcal{R}}(\pi, \hat{P}^{\boldsymbol{\alpha}^+}_{\mathbf{x}^+}, \hat{Q}^{\boldsymbol{\alpha}^-}_{\mathbf{x}^-})$$

- With the linear “loss” function, reversing the order of the two approximations would not work, and is thus not equivalent to the ERM inductive principle.

Conclusions

- Two views of binary experiments: "Generative" and "Discriminative"
- One-to-one correspondence: two views of the same underlying situation
- Parametrisation via weight functions helps (details omitted; see paper)
- Suggests a complementary viewpoint from which to derive MMD and SVM

Four Postdoctoral / Faculty positions available at ANU / NICTA!