Discriminative and Generative Views of Binary Experiments

Robert Williamson

December 2009

Joint work with Mark Reid
Consider the classical result

\[ L^{0-1}(\frac{1}{2}, P, Q) = \frac{1}{2} - \frac{1}{4} V(P, Q) \]  \hspace{1cm} (1) \]

where

\[ L^{0-1}(\frac{1}{2}, P, Q) = \inf_{r \in \{0,1\}^{X}} \mathbb{E}(X,Y) \sim \mathbb{P}[\ell^{0-1}(r(X), Y)]. \]

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

\[ M = \pi P + (1 - \pi)Q \quad \text{and} \quad \eta = \pi \frac{dP}{dM} \]
A function $\hat{\eta} : \mathcal{X} \rightarrow [0, 1]$ 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

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

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 $\eta$-average of the point-wise loss for $\hat{\eta}$:

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

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

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

- $\ell$, $L$ and $\mathbb{L}$ denote loss, point-wise and full risk of $\hat{\eta}$:
- The combination of a loss $\ell$ 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:

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

where

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

is the point-wise Bayes risk.
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 \( \ell \) satisfies \( L(\eta) = L(\eta, \eta) \) 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:

\[
L(\eta, \hat{\eta}) = L(\hat{\eta}) + (\eta - \hat{\eta})L'(\hat{\eta})
\]
Definition

The $f$-divergence of $P$ from $Q$ is

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

where $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ is convex and $f(1) = 0$. 
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Divergence Name</th>
<th>$f(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{V}$</td>
<td>Variational</td>
<td>$</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Liebler</td>
<td>$t \ln t$</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>Triangular Discrimination</td>
<td>$(t - 1)^2 / (t + 1)$</td>
</tr>
<tr>
<td>I</td>
<td>Jensen-Shannon</td>
<td>$\frac{1}{2} t \ln t - \frac{(t+1)}{2} \ln(t + 1) + \ln 2$</td>
</tr>
<tr>
<td>T</td>
<td>Arithmetic-Geometric Mean</td>
<td>$\left( \frac{t+1}{2} \right) \ln \left( \frac{t+1}{2\sqrt{t}} \right)$</td>
</tr>
<tr>
<td>J</td>
<td>Jeffreys</td>
<td>$(t - 1) \ln(t)$</td>
</tr>
<tr>
<td>$h^2$</td>
<td>Hellinger</td>
<td>$(\sqrt{t} - 1)^2$</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>Pearson $\chi$-squared</td>
<td>$(t - 1)^2$</td>
</tr>
<tr>
<td>$\Psi$</td>
<td>Symmetric $\chi$-squared</td>
<td>$\frac{(t-1)^2(t+1)}{t}$</td>
</tr>
</tbody>
</table>
Theorem

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

$$\phi(c) := \frac{1 - c}{1 - \pi} f(\lambda_\pi(c)), \quad \mathbb{L}(c) := -\phi(c)$$

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

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

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 $\phi$ in terms of a chosen $f$ such that the $f$ divergence $\mathbb{I}_f(P, Q)$ between $P$ and $Q$ equals the statistical information $\Delta \mathbb{L}(\eta, M)$ which equals the generative Bregman divergence $\mathbb{B}_\phi(\eta, M)$. 
The statistical information is the difference between the prior and posterior Bayes risk:

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

The generative Bregman divergence is

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

where $B_\phi$ is a standard Bregman divergence with respect to the convex function $\phi$. 

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]
Consider the following generalisation of $V$:

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

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 $\mathbb{L}^{\text{lin}}_{\mathcal{R}}(\pi, P, Q)$

**Theorem**

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

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

and the $r$ that attains $\mathbb{L}^{\text{lin}}_{\mathcal{R}}(\pi, P, Q)$ 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* $\phi : \mathcal{X} \rightarrow \mathcal{H}$ 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

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

- Thus

  $$\mathbb{L}_{\mathcal{R}}^{\text{lin}}(\pi, P, Q) = 1 - \frac{1}{4} \| \mathbb{E}_P \phi - \mathbb{E}_Q \phi \|_{\mathcal{H}}.$$
Given an independent identically distributed sample 
\( w = (w_1, \ldots, w_m) \in X^m \) the \( \alpha \)-weighted empirical distribution 
\( \hat{P}_w^\alpha \) with respect to \( w \) is defined by

\[
d \hat{P}_w^\alpha := \sum_{i=1}^{m} \alpha_i \delta(\cdot - w_i)
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_m) \), \( \alpha_i \geq 0 \), \( i = 1, \ldots, m \) and \( \sum_{i=1}^{m} \alpha_i = 1 \).

We will write \( \hat{E}_w^\alpha \phi := \hat{E}_{\hat{P}_w^\alpha} \phi = \sum_{i=1}^{m} \alpha_i \phi(w_i) \).

Thus

\[
V_{\mathcal{R}, \frac{1}{2}}^2(\hat{P}_w^\alpha, \hat{P}_z^\beta) = \frac{1}{2} \| \hat{E}_w^\alpha \phi - \hat{E}_z^\beta \|_{\mathcal{H}}^2.
\]
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 $\mathbf{\alpha} = (\alpha_1, \ldots, \alpha_m)$.
- Thus

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

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

$$\sum_{i \in I} \alpha_i y_i = \frac{m^+ - m^-}{m}.$$
We have

\[2V_{B_{\mathcal{H}, \frac{1}{2}}}(\hat{P}, \hat{Q}) = \left\langle \hat{E}_P \phi - \hat{E}_Q \phi, \hat{E}_P \phi - \hat{E}_Q \phi \right\rangle\]

\[= \left\langle \sum_{i \in I^+} \alpha_i \phi(x_i) - \sum_{i \in I^-} \alpha_i \phi(x_i), \sum_{j \in I^+} \alpha_j \phi(x_j) - \sum_{j \in I^-} \alpha_j \right\rangle\]

\[= \left\langle \sum_{i \in I} \alpha_i y_i \phi(x_i), \sum_{j \in I} \alpha_j y_j \phi(x_j) \right\rangle\]

\[= \sum_{i \in I} \sum_{j \in I} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle\]

\[= \sum_{i \in I} \sum_{j \in I} \alpha_i \alpha_j y_i y_j k(x_i, x_j) =: J(\alpha, x).\] (2)

We now consider three different choices of \(\alpha\).
If we set $\alpha_i = \frac{1}{m}$, $i = 1, \ldots, m$, then (2) becomes

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

where $x^+ := (x_i)_{i \in l^+}$, $x^- := (x_i)_{i \in l^-}$.

$\text{MMD}_b$ 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 $\alpha$.
- Minimizing $J(\alpha, x)$ over $\alpha$ will maximize $\mathbb{L}^{\text{lin}}$ and is thus the most pessimistic choice:

$$
\min_{\alpha} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (3)
$$

s.t. $\alpha_i \geq 0, \quad i = 1, \ldots, m$ \quad (4)

$$
\sum_{i=1}^{m} \alpha_i y_i = \frac{m^+ - m^-}{m} \quad (5)
$$

$$
\sum_{i=1}^{m} \alpha_i = 1 \quad (6)
$$

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

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

where $\nu \in (0, 1]$ is an adjustable parameter.

- $\nu$ controls the sparsity of $\alpha$ 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 $\nu$-SVM algorithm.
The traditional Empirical Risk Minimization principle replaces $(P, Q)$ with $(\hat{P}_{x^+}, \hat{Q}_{x^-})$ 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.

Set $\alpha^+ = (\alpha_i)_{i \in I^+}$ and $\alpha^- = (\alpha_i)_{i \in I^-}$.

The derivation above corresponds to

```
\mathbb{L}(\pi, P, Q) \xrightarrow{\text{Empirical Approximation (uniform)}} \mathbb{L}(\pi, \hat{P}_{x^+}, \hat{Q}_{x^-}) \xrightarrow{\text{Restrict Class}} \mathbb{L}_R(\pi, \hat{P}_{x^+}, \hat{Q}_{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!