Degrees of supervision

Darío García-García and Robert C. Williamson
Research School of Information Science and Engineering
The Australian National University
{dario.garcia,bob.williamson}@anu.edu.au

Abstract

Many machine learning problems can be interpreted as differing just in the level of supervision provided to the learning process. In this work we provide a unifying way of dealing with these different degrees of supervision. We show how the framework developed to accommodate this vision can deal with the continuum between classification and clustering, while also naturally accommodating less standard settings such as learning from label proportions, multiple instance learning,... All this emanates from a simple common principle: when in doubt, assume the simplest possible classification problem on the data.

1 Introduction

Can we think of clustering as unsupervised classification? Should we consider classification a supervised version of clustering? Is it possible to employ the same machinery for solving both problems? These questions hint at the interest of having a unified way of looking at machine learning problems with different *degrees of supervision*. In this paper we present a framework that can naturally handle the continuum of supervision degrees, with clustering and fully supervised classification as two extremes of this continuum, which also encompasses less well-known problems. Several examples of problems that can be cast together under our framework include

- **Supervised classification:** In this case the supervision information is comprised of *hard* labels $y \in Y = \{1, \ldots, K\}$ for points in a set called the training set.
- **Supervised classification with noisy labels:** In this scenario the labels can be contaminated with noise (or equivalently, we are not totally sure of them).
- **Semi-supervised learning:** In a semi-supervised setting we are given both a labeled and an unlabeled set of points, all of which are assumed to come from the same underlying marginal distribution (Chapelle et al., 2006).
- **Multiple Instance Learning:** The supervision information in this case is comprised of bags of points, each one of those bags being labeled as negative (if it contains no positive instances) or positive (if it contains at least one positive instance) (Dietterich et al., 1997).
- **Label proportions:** Here the supervision is given in the form of prior class probabilities P_i over bags of points $S_1, \ldots, S_M$ (Quadrianto et al., 2009).
- **Clustering (with balance penalties):** In a clustering problem there is no label information to guide the learning process, the only input is a set of points. In some cases, there is some prior knowledge about the size of the clusters, which is encoded using some balance constraints or penalties.
- **Partial labels:** In this scenario each sample can be associated to a set of several candidate labels. The candidate pool always includes the correct label and it may also include an ambiguity set Cour et al. (2011).

In addition to this, learning problems can appear in both inductive and transductive scenarios. In a transductive setting, the supervision information includes the set of test points where we are going to evaluate our algorithm. This way, there is no need to learn a predictor: we only need the actual predictions at the test points. In contrast, in an inductive setting there is no such knowledge about test time, so we must provide a means of obtaining predictions at arbitrary points, i.e. a predictor function. Note that this provides a way of drawing a distinction between the usual concept of clustering (i.e. find a sensible labeling for a bunch of unlabeled points) from that of unsupervised classification. The standard clustering problem can be understood as a transductive learning task, since we know from the beginning the set of points that we are ever going to predict on. In contrast, if the desired output is a predictor that can be applied out-of-sample, then we are in an inductive setting. In this sense, unsupervised classification conceptually corresponds with a smaller degree of supervision than that of standard (transductive) clustering.

To get a grasp of our framework, start by considering the following vision of a clustering task: we want to infer the most-informative classification experiment that we can pose on the data when we restrict ourselves to "smooth" statistics[1]. The posterior probability η is a sufficient statistic for such a statistical experiment, so we can think of finding a highly informative posterior function having some restrictions/penalties on its smoothness. This concept can be generally applied as a general principle for learning with incomplete supervision. By this we mean any case where there is not a fixed and unique label/target for each point in the dataset.

Given a sample $X = \{x_1, \ldots, x_N\}$ of points $x_i \in \mathcal{X}$ we can define the general shape of our objective functions for learning as follows

$$T_X(\eta) = \mathcal{D}_X(\mathcal{S}, \eta) + \mathcal{R}_X(\eta) \tag{1}$$

where $\mathcal{S}$ is the supervision information (to be defined more precisely later), $\mathcal{D}_X$ is a functional measuring discriminative power / information of the statistical experiment corresponding to the posterior probability η and $\mathcal{R}_X$ measures smoothness. Both functionals are *data-dependent*. This is analogous to the standard loss + regularization objective function which is standard machine learning. Our contribution will be mainly focused on showing how such a simple expression can encompass a whole range of learning problems. Let us explain the meaning of each one of its terms:

- **Smoothness**: Thinking of clustering in the above terms is naturally ill-defined if we allow the posterior probability function (or, in general, a sufficient statistic) to be any measurable function. Intuitively, any labeling is itself a maximally informative measurable function, so any labeling will be a perfectly valid solution to our clustering problem. It is then necessary to restrict the statistics to a smooth class of functions (or, dually, impose some penalty enforcing that smoothness). The requirement of a smoothness term should not be seen as a weakness of our interpretation: smoothness is the key behind almost any imaginable clustering algorithm. In the following section we show how maximum likelihood learning of mixture models (e.g. Gaussian mixture models) intrinsically implies a certain notion of smoothness. Another clear example is Spectral Clustering (SC), which can be interpreted as finding the smoothest zero-mean, unit-energy function on the data points, where the smoothness is measured in terms of the Laplacian operator (von Luxbürg, 2007). Spectral Clustering also shows the need for maximizing some notion of the information in the solution function: there is no guarantee that the function returned by spectral clustering will be in any sense close to a proper clustering function. By this we mean a function which takes values on a discrete set $Y = \{1, \ldots, K\}$ corresponding with the different clusters. From the point of view of classification, different notions of smoothness is related to regularization and generalization.
- **Information**: The *information* in a statistical experiment can be measured in terms of the dispersion of the likelihood ratio or the posterior probability functions (Ginebra, 2007), which are sufficient statistics. Specifically, a highly informative experiment is one whose sufficient statistic deviates a lots from its mean. In the limit case, a totally informative experiment will present a posterior probability function supported just on the vertices of the

[1] In this paper we use the word smooth to refer to a concept more general than the analytical meaning of continuous derivatives. In fact, we consider smoothness as a subjective measure of the simplicity of a function. For example, a fairly standard notion of smoothness for functions in $\mathbb{R}$ is related to the squared of the second derivative of the function. This way, sharp bends are penalized.

simplex, while a totally uninformative experiment will correspond with a posterior lying on the center of the simplex. Dispersion is measured using convex functionals (usually, expectations of convex functions). The other side of the coin is the *uncertainty*, which can be intuitively thought of as lack of information. Arguably the best known measure of uncertainty is Shannon's entropy (Cover and Thomas, 1991). However, it is possible to think of many alternative measures of entropy by relaxing the original axiomatic definition of Shannon and thinking in terms of other desirable properties (Ginebra, 2007). Our emphasis in this paper will be on this part of the learning objective function. As we will see, if our measure of discriminative power is based on *proper losses* (Reid and Williamson, 2011), then learning in our framework can be understood as following a generalized min-entropy principle.

The following sections develop these ideas. As an appetizer, in Section 2 we start by casting Maximum Likelihood and Classification Maximum Likelihood methods under our framework. Section 3 presents one of the key points of the paper, the interpretation of point-wise risks of proper losses as cross-entropies and their use for defining information functionals. Then, in Section 4 we introduce a general way of considering supervision information for classification problems, and how to handle that information. Section 6 is devoted to a small discussion of inductive and transductive learning and how they translate into our frameworks. Finally, Section 7 looks back on the main points of the paper.

2 Maximum likelihood (ML) and Classification Maximum Likelihood (CML) for Clustering

In this section we will see how Maximum Likelihood (ML) estimation of a mixture probabilistic model involves maximizing a certain (potentially degenerate[2]) notion of smoothness. We also show how another inference principle for mixture models, namely Classification Maximum Likelihood (CML) (Scott and Symons, 1971), intrinsically enforces both smoothness and informative posteriors. In spite of its name, CML is used for clustering purposes.

2.1 Maximum Likelihood (ML)

Given a parametric family of probability distributions $\{P_\theta\}_\Theta$, the ML estimate for θ based on the sample X is given by

$$\begin{aligned}\hat{\theta}_{\text{ML}} &= \arg\max_\theta P_\theta(X) = \arg\max_\theta \log P_\theta(X) = \arg\max_\theta \sum_{i=1}^N \log P_\theta(x_i) \\ &= \arg\min_\theta \sum_{i=1}^N \log \frac{1}{P_\theta(x_i)} \end{aligned} \tag{2}$$

In the asymptotic case we have

$$\hat{\theta}_{\text{ML}}(X) = \arg\min_\theta \mathbb{E}_{x\sim P}\left[\log \frac{1}{P_\theta(x)}\right] \tag{3}$$

so ML estimation corresponds with minimizing the cross-entropy between the generating distribution P and the parametric approximation P_θ. Equivalently, we can write

$$\hat{\theta}_{\text{ML}} = \arg\min_\theta \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{P_\theta(x)}\right] = \text{KL}(P||P_\theta), \tag{4}$$

where KL stands for the Kullback-Leibler divergence. We can think of ML as maximizing a notion of smoothness given by the likelihood of the parameters under the data. For clustering purposes, once a mixture model is learned via ML each point is assigned to the cluster corresponding to the mixture component with a higher posterior probability. However, this objective function does

[2]The usual problem with ML estimation of mixture models if the mixture components can have arbitrarily small variance.

not explicitly enforce good separation between clusters, in the sense that the probabilities that two different components assign to some samples can be very close (or equal) without any penalty. This can yield posterior probabilities which are quite "flat" and thus clusters with a high degree of overlapping.

2.2 Classification Maximum Likelihood (CML)

The Classification Maximum Likelihood paradigm can be traced back to Scott and Symons (1971), but in the following we take a different point of view. Assuming that the different classes/components are a-priori equally likely, we can write:

$$\begin{aligned}
\hat{\theta}_{\text{CML}}(X) &= \arg\max_{\theta,Y} P_\theta(X|Y) = \arg\max_{\theta,Y} \log P_\theta(X|Y) = \arg\max_{\theta,Y} \sum_{i=1}^{N} \log P_\theta(x_i|y_i) \\
&= \arg\max_{\theta,Y} \sum_{i=1}^{N} \log P_\theta(y_i|x_i) + \log P_\theta(x_i) \\
&= \arg\min_{\theta,Y} \sum_{i=1}^{N} \log \frac{1}{P_\theta(y_i|x_i)} + \log \frac{1}{P_\theta(x_i)} \\
&= \arg\min_{\theta,Y} \underbrace{\sum_{i=1}^{N} l_{\log}(y_i, \eta_\theta(x_i))}_{\mathcal{D}_X^{\text{CML}}(\eta_\theta)} + \underbrace{\log \frac{1}{P_\theta(x_i)}}_{\mathcal{R}_X^{\text{CML}}(P_\theta)}
\end{aligned} \tag{5}$$

This way it is readily seen that the CML principle implies both smoothness of the posterior probability estimate (as controlled by $\mathcal{R}_X^{\text{CML}}$) and discrimination performance (as controlled by $\mathcal{D}_X^{\text{CML}}$). The smoothness is once again measured in terms of the marginal probability of the data given the model that generates the posterior probability, so it is not just a functional of η_θ. That is because so far we are in a *generative* framework. Alternatively we can stay in a purely *discriminative* framework if we define notions of smoothness that depend just on η.

3 Cross-Entropies and proper losses

In this section we show how the concept of cross-entropy is closely related to discriminative learning, and how it can be naturally generalized using the class of proper losses. In a nutshell, minimizing a cross-entropy measure yields optimal decisions according to the loss at hand.

Let us focus on the information term of the r.h.s. of Eq. (5), corresponding to the log-loss between the estimated posterior and a optimal (based on that posterior probability) cluster assignment. This term can also be understood as follows

$$\begin{aligned}
\mathcal{D}_X^{\text{CML}} &= \min_{\theta,Y} \sum_{i=1}^{N} \log \frac{1}{P_\theta(y_i|x_i)} = \min_{\eta\in\mathcal{F}_\Theta} \sum_{i=1}^{N} \min_{y_i} \sum_{j=1}^{K} I(y_i = j) \log \frac{1}{\eta_j(x_i)} \\
&= \min_{\eta\in\mathcal{F}_\Theta} \sum_{i=1}^{N} \min_{p\in\tilde{\Delta}^K} H(p, \eta(x_i)),
\end{aligned} \tag{6}$$

where $H(p,q)$ is the *cross-entropy* between distributions p and q, and $\tilde{\Delta}^K = \{e_1, \ldots, e_K\}$ stands for the set of vertices of the $K-$simplex. We have changed the support of the optimization to explicate the fact that a probabilistic model defines a class $\mathcal{F}_\Theta$ of posterior probability functions. For example, a mixture model consisting of isotropic Gaussians of equal covariance matrices implies a class of posterior probability functions which are given by logistic functions with some constraints on the parameters. Those constraints relate the parameters with the means and variances of the Gaussians. Then, the role of the inference principle determines how we choose a function amongst that class.

The cross-entropy can be alternatively written as

$$H(p,\eta) = H(p) + KL(p||\eta). \tag{7}$$

When p is a vertex of the probability simplex, then $H(p) = 0$ and the expression reduces to $H(p, \eta) = KL(p||\eta)$.

Abusing the notation we will refer by η to both the posterior probability function and its evaluation at a certain point. The use should be clear by the context. It is natural to think about what happens if we substitute the log-loss for other loss functions. For reasons that will become clear later, we consider the subset of *proper losses* l_ϕ (Reid and Williamson, 2011), which are closely related to the well-known family of Bregman divergences B_ϕ. They can be defined in terms of a convex function $\phi : \Delta^K \to \mathbb{R}$ as

$$l_\phi(y, \eta) = B_\phi(e_y, \eta) = \phi(e_y) - \phi(\eta) - \langle e_y - \eta, \nabla\phi(\eta)\rangle, \tag{8}$$

where $y \in Y$ is a label and e_y is the vertex of the simplex corresponding to that label. We can then write

$$\min_{p\in\tilde{\Delta}^K} \mathbb{E}_{y\sim p} l_\phi(y, \eta_y) = \min_{p\in\tilde{\Delta}^K} J_\phi(p) + B_\phi(p, \eta) \equiv \min_{p\in\tilde{\Delta}^K} H_\phi(p, \eta), \tag{9}$$

where $J_\phi(p)$ is the Jensen gap functional $J_\phi(p) = \mathbb{E}_{y\sim p}[\phi(y)] - \phi(\bar{p})]$ (with $\bar{p} = \mathbb{E}[p]$) and B_ϕ is the ϕ-regret or Bregman divergence associated to the convex function ϕ. This expression provides a nice generalization of cross-entropy, since it is an additive expression consisting of an "entropy" term (see Sec. 3.1) and an approximation error: the entropy is generalized from Shannon entropy to the class of Jensen gap functionals, and the approximation error from Kullback-Leibler to the family of Bregman divergences. Note that when $C = \tilde{\Delta}^K$ there is no uncertainty in p and $J_\phi(p) = 0$.

We can relax the assumption that $p \in \tilde{\Delta}^K$ for $p \in C \subseteq \Delta^K$, so that the information functional will be defined in terms of

$$\min_{p\in C\subseteq\Delta^K} H_\phi(p, \eta) \tag{10}$$

Note that usually minimization of proper losses is carried out in the second argument. In fact, these losses are intrinsically characterized by the following property

$$\arg\min_{p\in\Delta^K} \mathbb{E}_{y\sim\eta}[B_\phi(\eta, p)] = \arg\min_{p\in\Delta^K} J_\phi(\eta) + B_\phi(\eta, p) = \eta, \tag{11}$$

with a minimum of $J_\phi(\eta)$. Here, however, we are minimizing in the first argument, so the concept is quite different.

Figure 3 shows how the regret and the Jensen gap relate to each other in the binary case. Denote by s_ϕ the segment connecting $\phi(0)$ and $\phi(1)$. Then draw a tangent of ϕ at η, which we can call t^η_ϕ. As can be seen from the figure, the value $H_\phi(p, \eta) = J_\phi(p) + B_\phi(p, \eta)$ is then given by the difference $s_\phi(p) - t^\eta_\phi(p)$, so

$$H_\phi(p, \eta) = [\phi(1) - \phi(0) - \phi'(\eta)]\, p + \phi'(\eta)\eta. \tag{12}$$

This is a linear function of p, so the optimum $p^* = \arg\min_p H(p, \eta)$ is obviously

$$p^* = \begin{cases} 0, & \phi'(\eta) < \phi(1) - \phi(0) \\ 1, & \phi'(\eta) > \phi(1) - \phi(0) \end{cases} \tag{13}$$

If $\phi'(\eta) = \phi(1) - \phi(0)$, then $H(p, \eta)$ is constant for every $p \in [0, 1]$. These results translate easily into the multiclass case, where we can write

$$p^* = e_k, \qquad k = \arg\min\left[\phi(\mathbf{e}) - \nabla\phi(\eta)\right], \tag{14}$$

where the $\arg\min$ should be interpreted in a vectorial sense (i.e. it returns the index of the largest component of the corresponding vector). If $\phi(e_0) = \phi(e_1) = \ldots = \phi(e_K)$, this expression simplifies to

$$p^* = e_k, \qquad k = \arg\max \nabla\phi(\eta). \tag{15}$$

The above expressions assume that the maxima are unique. Whenever $\phi(\mathbf{e}) - \nabla\phi(\eta)$ presents several maxima, then p^* need not necessarily be a vertex of the probability simplex, but it can also belong to the portion of the boundary given by convex combinations of the indices in the $\arg\min$.

Conceptually, this result states that *minimizing the generalized cross-entropy $H_\phi(p, \eta)$ over p amounts to finding the optimal "hard" classification corresponding to η, which can be decided just*

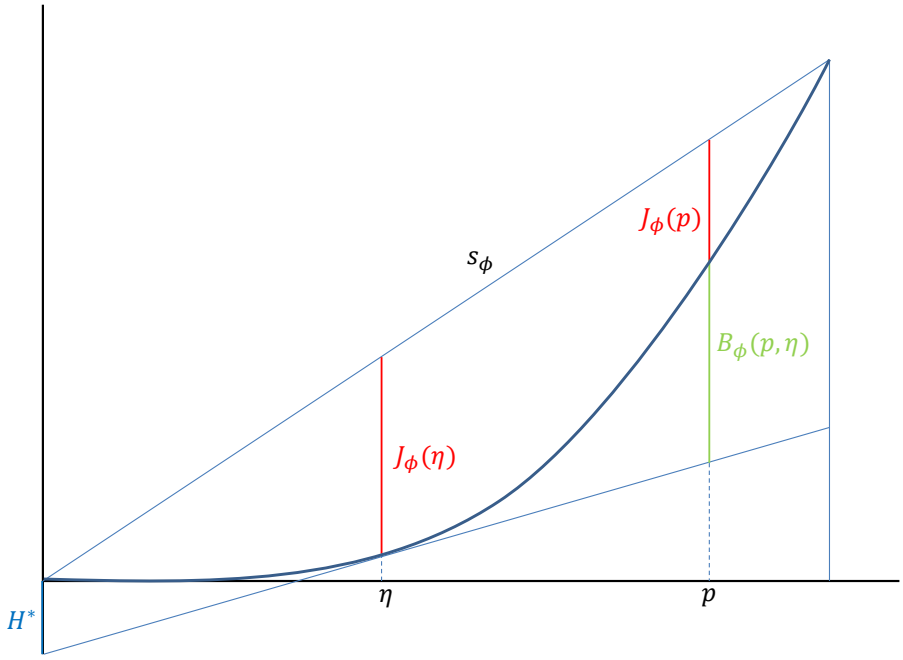

Figure 1: Regret and Jensen's gap

in terms of the value of the gradient $\nabla\phi(\eta)$. This way, we see that there is no benefit in increasing C in Eq. (10)[3] since

$$\min_{p\in\tilde{\Delta}^K} H_\phi(p,\eta) = \min_{p\in C} H_\phi(p,\eta), \quad \tilde{\Delta}^K \subseteq C \tag{16}$$

Recall that this result directly applies to Shannon's cross-entropy, since it is a special case of our generalized cross-entropies. This arises from the fact that the log-loss is a Bregman divergence (Reid and Williamson, 2011).

Using this result and the structure of $H_\phi(p,\eta)$, we can also write

$$\arg\min_{p\in\Delta^K} H_\phi(p,\eta) = \arg\min_{p\in\tilde{\Delta}^K} B_\phi(p,\eta) = \arg\min_k B_\phi(e_k,\eta) \equiv B^u_\phi(\eta) \tag{17}$$

and define, with some abuse of notation

$$\mathcal{D}_{X,\phi}(\eta) \equiv \frac{1}{N}\sum_{i=1}^{N} B^u_\phi(\eta(x_i)), \tag{18}$$

which can be interpreted as a generalized measure of the entropy of η. So a clustering algorithm minimizing the objective function

$$T_X(\eta) = \mathcal{D}_{X,\phi}(\eta) + \mathcal{R}_X(\eta) \tag{19}$$

achieves a trade-off between the smoothness of the posterior probability estimate (as measured by $\mathcal{R}$) and a notion of distance to a labeling function (i.e. generalized entropy) measured by the ϕ-regret B_ϕ.

Many semi-supervised learning methods try to minimize the entropy $H(\eta)$ of the estimated posterior probability on unalebeld points Grandvalet and Bengio (2004); Cid-Sueiro and a. Figueiras-Vidal

[3]Note that this is no longer true when a regularization term is introduced

(2001); Chapelle et al. (2006). Since $\min_{p\in\tilde{\Delta}^K} H_\phi(p,\eta) \leq H_(\eta,\eta) = H(\eta)$, those methods can be understood as either minimizing an upper bound of the optimal cross-entropy or as "imputing" the estimated posterior probability as the target for unlabeled points.

3.1 Recap: What are we saying about entropies?

We have assimilated Jensen gaps with entropies, which is justified mainly by the following points

- $J_\phi(p) = 0 \quad \forall p \in \tilde{\Delta}^K$
- $J_\phi(p)$ is concave. This is easy to see: let $m = \alpha p + (1-\alpha)q$. Then

  $J_\phi(m) = \mathbb{E}_{x\sim m}[\phi(x)] - \phi(\bar{m}) = \alpha\mathbb{E}_{x\sim p}[\phi(x)] + (1-\alpha)\mathbb{E}_{x\sim q}[\phi(x)] - \phi(\alpha\bar{p} + (1-\alpha)\bar{q})$,

  where $\bar{p}$ denotes the mean of distribution p. Since ϕ is convex, $\phi(\alpha\bar{p} + (1-\alpha)\bar{q}) \leq \alpha\phi(\bar{p}) + (1-\alpha)\phi(\bar{q})$ and so $J_\phi(m) \geq \alpha J_\phi(p) + (1-\alpha)J_\phi(q)$

We can constrain the class Φ of convex functions so that we get other interesting properties if $\phi \in \Phi$. Specifically, we want to enforce that $J_\phi(p)$ is maximized when p is non-informative, i.e., it lies on the center of the simplex. We can write the generalized entropy $H_\phi(p)$ in vector form

$$H_\phi(p) = \sum_{j=1}^{K} p_k\phi(e_k) - \phi(p) = p^T\phi(\mathbf{e}) - \phi(p),$$

where $\mathbf{e} = [e_1, \ldots, e_k]^T$. The natural gradient of such a function is given by

$$\nabla H_\phi(p) = \phi(\mathbf{e}) - \nabla\phi(p). \tag{20}$$

Recall the well-known conditions for maximizers of concave functions over the probability simplex (e.g. Th. 4.4.1 in Gallager (1968)). A concave function f achieves its maximum at a point $p = [p_1, \ldots, p_K] \in \Delta^K$ iff for any $c \in \mathbb{R}$

$$\frac{\partial f(p)}{\partial p_i} = c \qquad \forall i : p_i > 0$$
$$\frac{\partial f(p)}{\partial p_i} \leq c \qquad \forall i : p_i = 0.$$

Then, if we want to enforce $H_\phi(p) \leq H_\phi([1/K, \ldots, 1/K]) = H_\phi(c_K)$ we just need the condition

$$\nabla\phi(c_K) = \phi(\mathbf{e}). \tag{21}$$

In the case where $\phi(e_k) = 0$ for all k, this amounts to saying that the gradient $\nabla\phi$ vanishes at the center of the simplex. Note that these properties are exactly the requirements for an *uncertainty* measure of Ginebra (2007).

4 The supervision continuum

Let us look back on Eq. (17), which assigns a notion of information to the posterior function $\eta(x_i)$ evaluated at each point x_i in the sample X. It does so by finding the optimal classification in terms of ϕ and η. In order to accommodate different degrees of supervision, our information measure should depend on that supervision. This way, we could say that we need a notion of *cross-information*. The generalized cross-entropies defined in the previous section are good candidates for this job. Imagine a semi-supervised scenario (Chapelle et al., 2006), where the supervision information $\mathcal{S}$ consists of labels corresponding to points in a subset $X_l \subseteq X$. Then, if $x \in X_l$ with a corresponding label $y \in \tilde{\Delta}^K$, we don't have to optimize over the possible labels and we directly have that the information for that labeled pair should be given by $B_\phi(y, \eta(x))$. Then, we could write the global measure of information for the semi-supervised case as

$$\mathcal{D}_{X,\phi}(\eta) \equiv \sum_{x_i\in X_l} B_\phi(y_i, \eta(x_i)) + \sum_{x_i\in X_u} B^u_\phi(\eta(x_i)) \tag{22}$$

4.1 Indefinite labelings

Alternatively, we can rewrite the above expression in the following manner, which explicitly shows how the terms for the labeled and unlabeled samples are intimately related:

$$\mathcal{D}_{X,\phi}(\mathcal{S},\eta) \equiv \frac{1}{N}\sum_{x_i}\min_{p\in P_i} H_\phi(p,\eta(x_i)), \tag{23}$$

where $P_i \subseteq \Delta^K$ is the set of *allowable states* for sample x_i. When there is no supervision information about x, then $\mathcal{P}(x) = \Delta^K$ (and thus the optimal p will lie in the vertices on the simplex, as shown above). This formulation also encompasses what we call the *indefinite label* scenario, where the label for a given point is not necessarily a vertex of the simplex, and not even a point in the simplex, but a *set of points in the simplex*. This is a way of introducing uncertainty in the labeling in a continuum that goes from total certainty (the label is a single point in the simplex) to total agnosticity (the label is the whole simplex). Under this conception, unlabeled data are assimilated to data with a totally agnostic labeling.

Note that this indefinite label scenario also encompass the case of *multiple labelers*. We may have a set of labelers, each one of them annotating a subset of samples. This way, some samples can be annotated by more than one labeler. The different labels would then form the set of allowable states for a given sample.

4.2 Standard classification

Trivially, the standard supervised classification setting is recovered as the limit case where $N_u = 0$ (or, equivalently, each sample has a unique allowed state, given by its label). This way, we see that our proposed framework naturally covers the whole range between clustering and supervised classification. The key point behind this is the use of Bregman divergences / proper losses between (given or optimal) labels and predictions.

4.3 The meaning of it all: Tracing the original assumption

Recall that the tagline was "assume the simplest possible classification problem on the data". We used this originally as a paradigm for clustering. Along the way, we applied it to unlabeled data in a semi-supervised setting, and we ended up using it for dealing with indefinite or uncertain labels. This ride has shown that this principle is inherently equivalent to a *generalized min-entropy* criterion when proper losses are used for measuring classification performance.

5 Generalizing the label information

So far we have considered the cases where the available information for learning purposes is given by labels of individual points. However, the situation can be more complex than that. Consider for example the *learning from label proportions* case (Quadrianto et al., 2009; Rueping, 2010), where the supervision information is given in the form of the number of samples from each class in certain bags of points.

In order to encompass those "non-standard" scenarios together with the notion of indefinite labels into our framework, we will start by defining a general form of supervision information. Out learning algorithm takes as inputs an (unlabeled) sample $X = X_u \cup X_l = \{x_1,\ldots,x_n\}$ and a set of labeled pairs $\{(S_i,P_i)\}_{i=1}^{N_s}$, where $S_i \in 2^X$ (i.e. the power-set of X) and $P_i \subseteq \Delta^K$. Then we can write the following general form for the information functional

$$\mathcal{D}_{X,\phi}(\mathcal{S},\eta) \equiv \frac{1}{|\mathcal{S}|}\sum_i \min_{p\in P_i} H_\phi\left(p, \frac{1}{N_i}\sum_{x_j\in S_i}\eta(x_j)\right). \tag{24}$$

Obviously, different weights can be assigned to the cross-entropy terms for different elements of $\mathcal{S}$. Here we ignore that option since it does not add much to the idea and could clutter the notation.

Let us go back to the list of learning problems at the beginning of the paper and check how they can be cast within this framework:

- **Supervised classification:** We recover this scenario if $X = X_l$, $S_i = \{x_i\}$ for all $i \in [1, N_l]$ and each $P_i \in \tilde{\Delta}^K$ and is a singleton.
- **Supervised classification with noisy labels:** In this case each P_i does not need to coincide with a certain vertex of the simplex, but can be a subset $P_i \subseteq \Delta^K$.
- **Semi-supervised learning:** The classical semi-supervised scenario is recovered if $S_i = \{x\}$ for all $x \in [1, N]$ and $P_i \in \tilde{\Delta}^K$ for $\{i : x_i \in X_l\}$ and $P_j = \Delta^K$ for $\{i : x_i \in X_u = X \setminus X_l\}$.
- **Multiple Instance Learning:** We can cast MIL under our framework by associating negative bags with $P_i = e_0$ (the vertex of the simplex corresponding to class 0) and positive bags with $P_i = \Delta^K \setminus e_0$.
- **Label proportions:** In this case each supervision pair is comprised of a bag of points $S_i \subseteq 2^X$, and the empirical probability conditional on the bag $P_i = \frac{1}{N_i} \sum_{x_i \in S_i} e_{y_i}$. Note also that the label proportions case encompasses the standard supervised case, which we can effectively recover when $P_i \in \tilde{\Delta}$ for all i.
- **Clustering (with balance penalties):** Let $X = X_u$ and $S_1 = X$. Then, the value of p_1 encodes a "prior" on the cluster size balance. The closer p_1 is to the center of the simplex, the more balanced we expect the clusters to be.
- **Partial labels:** In this case $S_i = \{x_i\}$ for all i and each $P_i \subseteq \tilde{\Delta}^K$. That is to say, each sample is associated with an element of the power set of the labels.

Obviously, those scenarios can be freely combined, showing the flexibility of the framework. Moreover, further generalizations can be easily achieved. For example, the ϕ function defining the loss/entropy/divergence measure can be made sample-dependent, i.e. we can have a set $\Phi = \{\phi_1, \ldots, \phi_N\}$. This can be useful, for example, for reflecting arbitrary cost structures.

5.1 Classification vs Probability Estimation

We can think of two different setups for a label-generating process: a deterministic and a stochastic one. In the deterministic setting, there is a certain function $f_D : \mathcal{X} \to \mathcal{Y}$ mapping the input space to the label space $\mathcal{Y} = \{1, \ldots, K\}$. In the stochastic setting, there is a function $f_S : \mathcal{X} \to \Delta^K$ mapping the input space to the space of probability distributions over the labels. Labels for a given sample are then obtained by sampling from that distribution. The deterministic setting can be then understood as a particular case of the stochastic case, when the range of f_S is restricted to the vertices of the simplex.

The deterministic setting yield tasks that can be (in theory) perfectly solved, that is, with zero Bayes risk, by finding the function f_D. However, in the stochastic setting there will be a non-zero Bayes risk given the inherent randomness of the labeling process. However, both setups can be made analogous by introducing the concept of *label noise*. We will consider for simplicity the binary case, although the reasoning is totally general. Assume that the labels $\hat{y}$ that we see are given by a noisy oracle which randomly flips the value of the true labels $y = f_D(x)$ with probability ϵ (i.e. a noisy-typewriter kind of channel), so that

$$Pr(\hat{y}|y) = (1 - \epsilon)I[\hat{y} = y] + \epsilon I[\hat{y} \neq y].$$

This label noise takes us from the deterministic setting to a stochastic setting. The key point is that we can consider departures from the deterministic behaviour (i.e. randomness) as intrinsinc (and thus, something relevant that we should identify) or extrinsic, due to noise (and thus something that we must adequately handle but not necessarily identify).

Analogously, when we are given a bunch of sample-label pairs (or more general supervision information) we can think of two kinds of tasks to solve: classification tasks, where the focus is to find good label assignments, and probability estimation tasks, where the goal is to estimate the conditional distribution of labels given the samples. Classification is usually linked to the 0-1 loss, while probability estimation is related to the family of proper losses. Naturally, pure classification is strongly connected to the deterministic setting while probability estimation relates more to the stochastic setting. If we solve a probability estimation task on a deterministic setting with label noise, we would be finding both the assignment function f_D and the noise ϵ. Since the noise

of the oracle is not something that need to identify for learning purposes, we would be solving a harder problem than the one we are actually concerned with (although it is easy to think of many applications where knowledge of that noise can be beneficial).

At this point it is clear that the choice between classification or probability estimation methods should be guided by the following consideration about our label generating process: is it a deterministic mapping corrupted with noise or is it an intrinsically stochastic mapping? That is to say, is the randomness in the labels intrinsic or extrinsic? Obviously that question needs to be answered in a per-case basis. The interesting point for us is that both paradigms can be naturally encompassed in our framework, using the notion of indefinite labels Let us put a simple example for this. Recall the partial label scenario, and imagine that we know that a certain sample x_i belongs either to class v or class j. Then, the corresponding allowable states are $P_i = \{v, j\}$. Alternatively, we may think of an stochastic version of this setting: observing labels v and j for sample x_i in a stochastic setting implies a set of allowable states $P_i = \{p \in \Delta^K : p = \alpha e_i + (1-\alpha)e_j\}$, which corresponds with the segment connecting th i^{th} and j^{th} vertices of the simplex.

Indefinite labels can also be used in conjunction with stochastic labels: for example, we may want to introduce the information that a certain sample in a binary problem has a posterior probability in the range $[a, b]$. This kind of supervision can also be included in our framework.

6 Inductive and Transductive settings

The standard classification scenario is an *inductive* one: our goal is to find a function which is able to correctly classify unseen examples. However, in a *transductive* scenario there is no need to find any function, since the test points are known during the learning phase, and the goal is just to correctly classify those points. Intuitively, we can think that we are interested just in the regression function evaluated at those points, instead of in the function itself. From this point of view, we can consider that a sample X defines an equivalence class on a function class $\mathcal{F}$, given by all the functions which produce the same results when evaluated on X. Alternatively, we can think of a class of functions on a discrete support given by the sample. Since the effect of the information term $\mathcal{D}_X$ is naturally restricted to points in X, the difference between inductive and transductive learning in our framework is hidden within the regularization term $\mathcal{R}_X$. For example, we could think of working with η in a Hilbert space, and using the RKHS norm as a regularization term. Then, we are working in an inductive setting, since we are considering the full functional form of η. Alternatively, we could use a Laplacian regularization functional (Belkin et al., 2006)

$$\mathcal{R}_X^L(\eta) = \sum_{i \neq j} w_{ij}(\eta(x_i) - \eta(x_j))^2 \tag{25}$$

Such a functional takes into account just the values $\eta(x)$ for $x \in X$. This way, the learning process would be purely transductive, since there is a total disregard about the behaviour of η outside the sample X. In fact, there is no need to even think of η as a function.

From our point of view, transduction and induction can also be thought of as different levels of supervision: in a transductive setting we know the points that we want to predict, so we can work directly with them. By contrast, in an inductive setting we are not given such information, so we need to provide a means of obtaining predictions for any arbitrary point.

7 Conclusions

Starting with a simple principle for clustering, which can be summarized as "find the easiest classification problem that can be posed on the data", we have built a framework that allows for defining classification objective functions for a huge set of scenarios in a unified fashion: clustering, semi-supervised, fully supervised, label proportions, multiple instance learning and many more. The key point is to define a flexible notion of *informativeness* of a posterior probability function that can naturally accommodate varyin g degrees of supervision. To this end, we have used the machinery of proper losses but interpreted in a new way. We have proposed that point-wise risks for proper losses are meaningful generalizations of Shannon's cross-entropy, given their natural decomposition as the sum of a (generalized) entropy plus an "error-term" expressed as a divergence measure. We

have studied their properties from this point of view, showing that the interpretation is meaningful. Finally, we have sketched how both inductive and transductive learning can take place under our general framework.

We believe that being able to view this whole range of problems from a common perspective opens up the possibilities of exploiting results and techniques developed for one problem and applying them to others.

Acknowledgments

DGG wants to thank Ulrike von Luxburg and Tiberio Caetano for helpful discussions.

References

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434.

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Cid-Sueiro, J. and a. Figueiras-Vidal (2001). On the structure of strict sense bayesian cost functions and its applications. *IEEE Transactions on Neural Networks*, 12(3):445–455.

Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience.

Dietterich, T., Lathrop, R., and Lozano-Perez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71.

Gallager, R. (1968). *Information theory and reliable communication*. Wiley.

Ginebra, J. (2007). On the measure of the information in a statistical experiment. *Bayesian Analysis*, 1(5):1–45.

Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing (NIPS) 17*.

Quadrianto, N., Smola, A., Caetano, T. S., and Quoc, V. L. (2009). Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374.

Reid, M. D. and Williamson, R. C. (2011). Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817.

Rueping, S. (2010). Svm classifier estimation from group probabilities. In *International Conference on Machine Learning*.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397.

von Luxbürg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4).