2.1 PAC learning with n noisy annotators

For any hypothesis $h \in \mathcal{C}$, the error rate (or generalization error) is defined to be the probability that $h(x) \neq c_i(x)$ for an instance $x \in \mathcal{X}$ that is randomly drawn according to $D$. The error rate of a hypothesis $h$ is given by $P^D(c_i \Delta h)$, where $c_i \Delta h \subseteq \mathcal{X}$ is the symmetric difference between sets $c_i$ and $h$, and $P^D(\cdot)$ is the probability of this event (calculated with respect to $D$). A hypothesis $h$ is said to be $\epsilon$-bad if its error rate is more than $\epsilon$, i.e. $P^D(c_i \Delta h) > \epsilon$. In the classical PAC model of [Valiant, 1984], the learner’s goal is to come up with a learning algorithm which outputs an $\epsilon$-bad hypothesis $h$ with probability at most $\delta$, where the probability is defined with respect to the distribution of training examples of a fixed size. Such a learning algorithm is known as PAC learning algorithm. In general, the error rate of the hypothesis chosen by a learning algorithm critically depends on the number of training examples supplied to the algorithm. Thus, a learning algorithm with single annotator is said to satisfy PAC bound with respect to the sample size $m(\epsilon, \delta)$ if the following condition holds true: $P^{m(\epsilon, \delta)}(P^D(c_i \Delta h) > \epsilon) < \delta$, where $h$ is the hypothesis output by the learning algorithm when trained on the $m(\epsilon, \delta)$ number of training examples. The sample size $m(\epsilon, \delta)$ is a non-negative integer valued function of the parameters $\epsilon$ and $\delta$. The probability $P^{m(\epsilon, \delta)}(\cdot)$ is taken over the distribution of $m(\epsilon, \delta)$ training examples (noisy or non-noisy). For a given algorithm, the smallest sample size $m^*(\epsilon, \delta)$ for which it still satisfies PAC bound is known as its sample complexity. Now, we extend the PAC learning framework to the case of $n$ noisy annotators; starting with the following definitions:

- **An instance** of the PAC learning problem is a set of specifications of instance space $\mathcal{X}$, concept class $\mathcal{C}$, true concept $c_i$, and sampling distribution $D$.

- **An annotation plan**, denoted by $m(\epsilon, \delta) = (m_1(\epsilon, \delta), \ldots, m_n(\epsilon, \delta))$, is a vector of sample sizes (number of examples) annotated by the $n$ annotators. This quantity is analogous to sample size $m(\epsilon, \delta)$ in the single annotator case. In rest of the paper, we use $m_i$ and $m_i(\epsilon, \delta), i = 1, 2, \ldots, n$, interchangeably.

- A learning algorithm for $n$ noisy annotators is said to satisfy PAC bound for annotation plan $m = (m_1, \ldots, m_n)$ if following holds true:

  $$P^{m_1, \ldots, m_n}(P^D(c_i \Delta h) > \epsilon) < \delta$$

  (1)

  Note that the noise rates $\eta_1, \ldots, \eta_n$ of the annotators could be very different. Hence the PAC bound depends not just on $\sum_{i=1}^n m_i$, but on the individual numbers $m_1, \ldots, m_n$ also. This motivates us to define the notions of feasible and infeasible annotation plans, as:

  - For a given learning algorithm, an annotation plan $m = (m_1, \ldots, m_n)$ is said to be **feasible** if the learning algorithm satisfies PAC bound (1) for every instance of the problem when training data is supplied as per this plan.

  - Given an algorithm, an annotation plan $m = (m_1, \ldots, m_n)$ is said to be **infeasible** if the algorithm fails to satisfy PAC bound (1) for at least one instance of the problem when training data is supplied as per this plan.

2.2 Feasible annotation plans for MDA

In this section, we consider a simple learning algorithm, namely Minimum Disagreement Algorithm (MDA) and derive PAC learnability bound on annotation plan complexity for this algorithm in the presence of $n$ noisy annotators. [Laird, 1988] analyzed this algorithm for single noisy annotator case. MDA outputs the hypothesis $h$, which minimizes the empirical loss, $L_e(h)$, on the training dataset. We describe MDA for multiple annotators, below.

**Algorithm 1 (MDA)** Let $\mathcal{D} = \{(x_{ij}^i, y_{ij}^i) \ i = 1, \ldots, n; j = 1, \ldots, m_i\}$ be the input training data, where $(x_{ij}^i, y_{ij}^i)$ is supplied by annotator $i$ in $j$th call. The empirical loss $L_e(h)$ for hypothesis $h$ is given as:

$$L_e(h) = \sum_{i=1}^n \sum_{j=1}^{m_i} 1(h(x_{ij}^i) \neq y_{ij}^i)$$

(2)

where $1(\cdot)$ is an indicator variable. Output hypothesis $h^* \in \mathcal{C}$, such that $L_e(h^*) \leq L_e(h), \forall h \in \mathcal{C}$ (use any tie breaking rule).