A computational teaching theory for Bayesian learners

Xiaojin Zhu

jerryzhu@cs.wisc.edu

In this talk, I will describe a new computational teaching theory. This new theory extends the original teaching dimension theory [1], incorporates the curriculum learning framework [2], and generalizes the special case analysis of [3]. The new theory may provide guidance and theoretical justification for teaching categorization tasks to human students.

The new theory has three entities: the task, the learner, and the teacher.

1. The task is a joint probability density p_{XY} over input X and label Y. Future test items will be drawn *iid* from p_{XY}. This is the same as standard learning theory for classification, except that the training data does not have to be drawn *iid* from it, as we will see next.

2. The learner is a Bayesian learner. The learner does not know p_{XY}. It operates over a parameter space Θ, which may not include the Bayes classifier under p_{XY}. It has a prior distribution $p(\theta)$ for $\theta \in \Theta$. It has a likelihood model $p(x, y \mid \theta)$. Given a training item (x_i, y_i), it will perform Bayesian update to form a posterior distribution over Θ. As in standard Bayesian learning, it can perform Bayesian updates sequentially over training items. It is agnostic to the *iid*-ness of the training items.

3. The teacher is clairvoyant. It knows p_{XY}, the learner's Θ, prior $p(\theta)$, and likelihood $p(x, y \mid \theta)$. However, it can only communicate with the learner by providing training items. There will be no outrageous collusion between the teacher and the learner because the learner can only perform Bayesian updates. The teacher's goal is to carefully "find" training items so the learner will be accurate.

More precisely, we want the risk of the learner $R_n \equiv \mathbb{E}_{p_{XY}}[\ell(p(y \mid x), y)]$ to be as small as possible after seeing a training set of size n, where ℓ is an appropriate loss function and $p(y \mid x)$ is the learner's predictive distribution under its posterior. Note that $R_n \geq R^*$, the minimum risk achievable by *any* distribution over Θ, and in turn $R^* \geq R^{Bayes}$ the Bayes risk of P_{XY}. It is reasonable to define the our goal as "finding" the smallest training set such that $R_n - R^* \leq \epsilon$.

How does the teacher "find" the training set? We need to further clarify the power of the teacher in two separate scenarios. In the first scenario, the teacher can *design* training items, i.e., teach with arbitrary (x, y) pairs that may have nothing to do with P_{XY}. This is arguably more powerful than the second scenario, where there is a fixed pool of items sampled *iid* from P_{XY} and the teacher must select training items from the pool. This separation is rather similar to that in active learning.

Under the above setup, I will show that teaching dimension and curriculum learning are two sides of the same coin. I will discuss a few human teaching experiments where some human teachers seem to follow the theoretical prediction on how they should behave.

References

[1] S. Goldman and M. Kearns. On the complexity of teaching. *Journal of Computer and Systems Sciences*, 50(1):20–31, 1995.

[2] Yoshua Bengio, Jérme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Léon Bottou and Michael Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, Montreal, June 2009. Omnipress.

[3] Faisal Khan, Xiaojin Zhu, and Bilge Mutlu. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems (NIPS) 25*. 2011.