Active learning with cautious oracle that provides relative answers

A1, A2

October 1, 2013

1 Basis model for active learning

Using the language of the 2010 survey by Settles.

Task: Train a classifier semi-supervised manner. We have a set $\mathcal{L}$ of labeled training samples, and we want to improve the classifier by querying an annotator for some more.

The scenario: Pool-based sampling. We have a pool $\mathcal{U}$ of unlabeled samples which we can use for querying the oracle for their label.

Query strategy: uncertainty sampling using entropy as query informativeness.

1.1 Cautious oracles

The annotator gives a label-probability vector instead of a single label (oracle) or random label (noisy oracle).

The idea: Annotator is modelled as the probability vector producer. We use the Dirichlet-distribution to model the annotator.

Model: Let $\mathcal{D} = \{(y, x)\}$ denote a set of items, where $y \in C$ with $|C| = m$ denote labels for the items, and $x \in \mathcal{X} \subseteq \mathbb{R}^p$ denote other features (covariates). Write for any x the correct label y_x.

Start with a training set of correcly labeled items $\mathcal{L} = \{(y_i; x_i)\}$. Denote $\mathcal{U} = \mathcal{D} \setminus \mathcal{L}$. The classifier model

$$\begin{aligned} y|x, \theta &\sim p(y|x, \theta) && (1) \\ \theta &\sim p(\theta) && (2) \end{aligned}$$

learns the θ from the $\mathcal{L}$. Active learning is about increasing the model fitness by querying most informative x to get y_x so as optimally increase upon $\mathcal{L}$ with minimum number of queries. The basic setting assumes an *oracle*, an external annotator which provides the correct label for any query $x^\star \in \mathcal{U}$ and which we then use to enhance the estimate of θ.

Now define a *cautious oracle*: when queried for a label for $x^\star \in \mathcal{U}$ the cautious oracle's answer is a distribution over C instead of the correct label:

$$\begin{aligned} h^\star &\sim p(h^\star|x^\star, \alpha), \quad h^\star \in \mathcal{H} = \{h \in \mathbb{R}_+ : 0 < \sum_{c \in C} h_c < \infty\} && (3) \\ p(y^\star = c'|x^\star) &= h_{c'} / \sum h_c && (4) \\ \alpha &\sim p(\alpha) && (5) \end{aligned}$$

We get the oracle from this by setting

$$p(h|x^\star, \alpha) = 1(h_{y_{x^\star}} > 0, h_c \neq 0 \ \forall \ c \neq y_{x^\star}).$$

In other words, the active learning maps queries with functions

$$\begin{aligned} \text{oracle} \quad f_o(x^\star) &= y_{x^\star} && (6) \\ \text{cautious oracle} \quad f_{co}(x^\star) &= h_{x^\star} && (7) \end{aligned}$$

How do we use the cautious oracle for active learning? The oracle is used in the following algorithmic manner:

1. Compute the informativeness $I(x) := I(x|\theta, \mathcal{L})$ of each $x \in \mathcal{U}$
2. Augment $\mathcal{L} := \mathcal{L} \cup \{(f_o(x^\star), x^\star)\}$ where $x^\star = argmax \ I(x)$.
3. Check some stopping rule.

The cautious oracle case suggest two approaches to the algorithm, of which both lead to the same probabilistic model. First is the 'heuristic' realisation approach, where we sample the label from the distribution provided by the oracle, viz.

2'. Augment $\mathcal{L} := \mathcal{L} \cup \{(\tilde{y}_{x^\star}, x^\star)\}$ where $x^\star = argmax \ I(x)$ and $\tilde{y}_{x^\star} \sim h_{x^\star}$.

This leads to a new source of noise in $\mathcal{L}$, which then is handled by the augmenting the classifier model.

The other option is to change the structure of the data. Let again $\mathcal{L}_t$ be the queried set by now consisting of pairs (h_x, x). Then

2". Augment $\mathcal{L} := \mathcal{L} \cup \{(h_{x^\star}, x^\star)\}$ where $x^\star = argmax \ I(x)$.

If $\mathcal{L}_0$ is the (clean) training set, and $\mathcal{L}_t$ is the queried set after t iterations, we add to the model (1)

$$\tilde{p}(h_x|x, \theta) = E_{\tilde{y}|h_x} p(\tilde{y}|x, \theta) \qquad (h_x, x) \in \mathcal{L}_t \tag{8}$$

This is then used for inferring θ, and for deriving new values for $I(x)$.

1.2 Example 1

Let the classifier be naive Bayes, i.e.

$$p(y = c|\theta, x) \propto p(c|\theta_c) \prod p(x_k|\theta_c)$$

where θ_c are the parameters corresponding to class c.

Let the informativeness measure be the entropy,

$$I(x) := E_{\theta|\mathcal{L}} \sum_c p(y = c|x, \theta) \log p(y = c|x, \theta).$$

2 Model for the cautious oracle

Can we infer

$$p(h|x, \alpha)?$$