Active learning
Thomas P. Minka

Abstract

This note defines the information in a new sample from a distribution, and uses it to determine the optimal queries to make now for good inferences later.

1 Introduction

Suppose we are receiving IID samples x from some distribution $p(x|\theta)$ where θ is unknown. Let D denote all the information that we have about θ, possibly including samples from $p(x|\theta)$. Then $p(\theta|D)$ quantifies our knowledge of θ so far.

Our goal is to learn the distribution $p(x|\theta)$ as well as possible, where "well" is quantified by the KL-divergence between our current best estimate, call it $q(x)$, and the true distribution. We don't know the true distribution exactly, but we do have a distribution about what it might be, so we can measure how well we are doing in terms of the expected KL-divergence:

$$\begin{aligned} I &= \int_\theta p(\theta|D)\mathcal{D}(p(x|\theta) \mid\mid q(x)) && (1) \\ &= \int_\theta p(x,\theta|D) \log \frac{p(x|\theta)}{q(x)} && (2) \end{aligned}$$

The first question is: what estimate $q(x)$ should we choose to minimize this quantity? By zeroing the derivative wrt $q(x)$, we find that

$$q(x) = \int_\theta p(x|\theta)p(\theta|D) = p(x|D) \quad (3)$$

That is, the best estimate is the mean density over x given by our prior knowledge.

If we take this as our estimate, then I can be interpreted as the "average divergence to the mean density" or the "variance about the mean density." We can rewrite it to obtain

$$\begin{aligned} I &= \mathcal{H}(x|D) - \mathcal{H}(x|\theta, D) && (4) \\ &= \mathcal{I}(x,\theta|D) && (5) \end{aligned}$$

the mutual information (Cover and Thomas) between x and θ, given our knowledge so far.

The next question is: what is the value of a new sample from the distribution? This depends on what criterion we are trying to optimize. If we are trying to learn as much as possible about

θ, then we want to decrease $\mathcal{H}(\theta)$ as much as possible. The expected decrease in entropy from observing a new sample is

$$\mathcal{H}(\theta|D) - \mathcal{H}(\theta|x, D) = \mathcal{I}(\theta, x|D) \tag{6}$$

That is, the value of a new sample is equivalent to our expected modeling error.

However, this objective is not appropriate for all tasks, because different θ's need not lead to significantly different distributions over x. It is possible for us to be very uncertain about θ and yet be very certain (in the sense of I) about $p(x|\theta)$.

An alternative is to measuring the decrease in $I(x, \theta|D)$ from a new sample. To compute this, we consider augmenting our information by x_0 to get $D' = D \cup x_0$. Then

$$\begin{aligned}
\mathcal{H}(x|\theta, D') &= \mathcal{H}(x|\theta, D) && (7)\\
\mathcal{H}(x|D') &= \mathcal{H}(x|x_0, D) && (8)\\
\mathcal{I}(x, \theta|D) - \mathcal{I}(x, \theta|D') &= \mathcal{H}(x|D) - \mathcal{H}(x|x_0, D) && (9)\\
&= \mathcal{I}(x, x_0|D) && (10)
\end{aligned}$$

the mutual information between x and x_0, two independent samples from the distribution. This makes sense in that it is only concerned with what one sample can tell us about a future sample, not what it can tell us about θ.

1.1 Classification

In classification, we want to model $p(c|x, \theta)$. Suppose we have prior knowledge about the form of $p(x|c, \theta)$ and some knowledge D about θ. Then we can classify a point x using the c that maximizes

$$p(c|x, D) = \frac{p(x|c, D)p(c|D)}{p(x|D)} \tag{11}$$

Our modeling error can be expressed as

$$\int_\theta p(\theta|D)\mathcal{D}(p(c|x, \theta) \mid\mid p(c|x, D)) = \mathcal{I}(c, \theta|x, D) \tag{12}$$

The question is: if we could ask for the true class c of one point x, which x should we choose?

By the preceding argument, if we want to learn the most about θ then we should pick the x_0 that maximizes $\mathcal{I}(c, \theta|x = x_0, D)$ or equivalently minimizes

$$\mathcal{H}(\theta|c, x = x_0, D) = \sum_{c_0} p(c = c_0|x = x_0, D)\mathcal{H}(\theta|(c_0, x_0), D) \tag{13}$$

This is the approach used by McCallum (1998), building on an earlier approximate scheme called query-by-committee. McCallum used sampling to approximate the integral over θ: an exact formula for multinomial classes is given in the next section.

If instead we want to learn the most about $p(c|x,\theta)$, i.e. give the best classification probabilities, then we should pick the x_0 that maximizes $I(c, c_0|x, D)$, where c_0 is the (unknown) class of x_0. This is equivalent to minimizing

$$\sum_{c_0} p(c = c_0|x = x_0, D)\mathcal{H}(c|x, (c_0, x_0), D) \tag{14}$$

Note that we are assuming in both cases that x_0 is synthesized by us or already included in D, so that x_0 itself does not carry new information about $p(c|x,\theta)$.

1.2 Regression

In regression, we want to learn the distribution $p(y|x,\theta)$ as well as possible. Suppose we already have some knowledge D about θ. We can choose to probe at any location x_0. Which one should we choose?

To learn the most about θ, we choose x_0 to minimize

$$\mathcal{H}(\theta|x = x_0, y, D) = \int_{y_0} p(y = y_0|x = x_0, D)\mathcal{H}(\theta|(x_0, y_0), D) \tag{15}$$

To learn the most about the value of y at point x, we choose x_0 to minimize

$$\int_{y_0} p(y = y_0|x = x_0, D)\mathcal{H}(y|x, (x_0, y_0), D) \tag{16}$$