Discriminative models, not discriminative training

Tom Minka
Microsoft Research Cambridge, October 17, 2005

Suppose you are given a dataset of pairs (x, c) where c is a class variable and x is a vector of features. Given a new x, you want to predict its class. The generative i.i.d. approach to this problem posits a model family

$$p(x, c \mid \theta) = p(x \mid c, \lambda)p(c \mid \pi) \tag{1}$$

and chooses the best parameters $\theta = \{\lambda, \pi\}$ by maximizing (or integrating over) the joint distribution (where D denotes the data):

$$p(D, \theta) = p(\theta) \prod_i p(x_i, c_i \mid \theta) = p(\theta) \prod_i p(x_i \mid c_i, \lambda)p(c_i \mid \pi) \tag{2}$$

Another approach, sometimes called "discriminative training" or "conditional training", chooses the best θ by maximizing (or integrating over) the conditional distribution:

$$p(C, \theta \mid X) = p(\theta) \prod_i p(c_i \mid x_i, \theta) \tag{3}$$

$$\text{where } p(c \mid x, \theta) = \frac{p(x, c \mid \theta)}{\sum_c p(x, c \mid \theta)} \tag{4}$$

While this is a valid way of obtaining a classifier, the description is misleading. To start with, the term "discriminative training" is a misnomer, because given a probabilistic model, there is only one correct likelihood and therefore only one correct way to train it. What is really going on in (3) is that the model has changed, not the training principle.

The correct way to derive (3) is to posit a new model family with an additional set of parameters θ':

$$q(x, c \mid \theta, \theta') = p(c \mid x, \theta)p(x \mid \theta') \tag{5}$$

$$\text{where } p(x \mid \theta') = \sum_c p(x, c \mid \theta') \tag{6}$$

Here $p(c \mid x, \theta)$ is the same as (4) and $p(x, c \mid \theta')$ is the same as (1) but with parameters θ'. The parameter sets θ and θ' have the same type but are independent. Now choose the best parameters (θ, θ') in the standard way by maximizing (or integrating over) the *joint* likelihood:

$$q(D, \theta, \theta') = p(\theta)p(\theta') \prod_i q(x_i, c_i \mid \theta, \theta') = p(\theta)p(\theta') \prod_i p(c_i \mid x_i, \theta)p(x_i \mid \theta') \tag{7}$$

Due to the model assumptions, the estimations of θ and θ' decouple, so the best θ is the same as in (3).

By taking this view, you have a consistent approach to statistical inference: you always model all variables, and you always use joint likelihood. The only thing that changes is the model.

You can also see clearly why discriminative training might work better than generative training. It must be because a model of the form (5) fits the data better than (1). In particular, (5) is necessarily more flexible than (1), because it removes the implicit constraint that $\theta = \theta'$. Removing constraints reduces the statistical bias, at the cost of greater parameter uncertainty.

Besides consistency and clarity, this view also has a practical advantage, in that you can easily blend between the generative and discriminative approach, e.g. to incorporate unlabeled data. All you do is use a prior $p(\theta, \theta')$ in which θ and θ' are coupled. The θ' parameter will adapt to fit the unlabeled x's, which then affects θ. By forcing the parameters to be equal, you recover the generative approach. With a softer coupling, you get discriminative semi-supervised learning.

To summarize: The term "discriminative training" should be abolished. Instead, we should refer to models of the form in (5) as *discriminative models*.

Acknowledgement. Martin Szummer and Chris Bishop helped clarify the presentation.