Forms of Learning — From Known Models and Unknown Losses to Unknown Models and Known Losses

Peter D. Grünwald

November 18, 2012

The accompanying table shows three modes of learning from i.i.d. data of the form

$$(x_1, y_1), \ldots, (x_N, y_N).$$

They are ordered from top to bottom in the order of *assuming less and less about the underlying mechanism generating the data, and assuming more and more about how the 'thing' (distribution, function, predictor) that is learned from data is to be tested.* They clearly show the special importance of the *log-loss* and the *squared loss* (mean squared error), which have sometimes been called "emperor" and "king" of all loss functions (0/1-loss is then "queen"). Log-loss and squared loss are used in Approaches 1 and 2 respectively, because, given enough data and a simple enough model, they give as output a $\hat{p}$ or $\hat{f}$ that, with high p^*-probability, is guaranteed to be actually a good approximation of the true p^* or f^*. Thus, log-loss and squared loss are meaningful during training even when, during the testing, other loss functions are to be used. They are not the only loss functions with this property, but they are definitely the ones for which this property holds most generally.

Most importantly, many of the learning models and methods (algorithms) that we describe during the lectures fit in *more than one* of these approaches, depending on what assumptions one is willing to make and how one will test one's data.

Approach I (the first row) is traditional statistical modeling based on assuming a *probability model* for the data. A few remarks about Approach I are in order:

1. In our notation we used a parametric model (with parameter vector θ), but the approach extends to nonparametric models as well.

2. Both 'frequentist' and 'Bayesian' statistical approaches fall in this row — although many Bayesian statisticians would not say that they 'assume the data are sampled from a true distribution', Yet Bayesian approaches still rely on modeling *all* aspects of a phenomenon *probabilistically*, and (usually) do not take into account the loss function L^* of interest during training — which makes them fall into Approach I.

3. Formally, any probability model for $\mathcal{X} \times \mathcal{Y}$ may be viewed as a 'generative' model, so mathematically the phrase 'generative model' is not very meaningful (unless in some special cases which we will not go into here). The term *generative* model is really about how we *think* of the model: we model the data as 'Nature first fixes some y (drawn from a distribution $p_\theta(Y)$). The value $Y = y$ chosen determines the distribution $p_\theta(X \mid y)$ and then, Nature draws X according to this distribution' (LDA and Naive Bayes are prototypical examples).

4. Any model $\mathcal{P}$ (set of distributions p_θ on $\mathcal{X} \times \mathcal{Y}$) may be changed into a corresponding discriminative model $\mathcal{P}'$ containing the corresponding conditional distributions $p_\theta(Y \mid X)$ for Y given X. Standard maximum likelihood and log-loss in $\mathcal{P}'$ correspond to conditional maximum likelihood and conditional log-loss in $\mathcal{P}$.

5. One may be inclined to think that one should *always* use discriminative models if the aim is to predict Y given X, but this is not always so, for at least two reasons: (a) the discriminative model may need more data before it outputs a good distribution than its generative counterpart; (b) in applications such as *semi-supervised learning*, if the model is indeed correct, one can use *unlabeled data* (i.e. data for which the Y-values are missing; such data is often much cheaper than complete data) to help guide the learning with generative models; but for discriminative models, unlabeled data are useless.

How do Our Algorithms Fit In?

Standard Least Squares Regression when $Y = \mathbb{R}$ fits into all three approaches!

Linear regression with least squares based on a linear model $f_\beta(x) = x^T\beta$ is the canonical algorithm within Approach II. However, if one is willing to assume *Gaussian noise with fixed variance* it also fits into Approach I. (then *least squares and maximum likelihood* become identical, and so do *squared loss and log-loss*). Moreover, if one is happy to use the algorithm, after learning, for the squared loss (i.e. one tests its performance by squared loss), it also fits into Approach III (note that least squares is ERM with the squared loss function).

Ridge Regression and Lasso fit into all three Approaches!
...for the same reason as above.

Least Squares Regression with linear model used for Classification ...surprisingly also fits into all three approaches.

To see how it fits into Approach III, note that we fit using least squares (i.e. by empirical risk minimization with the squared loss), but we intend to learn something about a different loss function — usually the 0/1-loss. So the squared loss is our 'proxy'.

Since we're doing least squares during training, this method automatically belongs to Approach II as well, at least *if* we are willing to assume that $f^* \in \mathcal{F}$, i.e. that our model contains the true function so that $Y = f^*(X)$ plus 0-mean noise.

To see how it fits into Approach III, take for simplicity the case where there are two classes, i.e. $\mathcal{Y} = \{0, 1\}$, so — as a simple calculation shows — $E_{P^*}[Y \mid X] = P^*(Y = 1 \mid X = x)$. Hence, under the assumption that $f^* \in \mathcal{F}$, we can also think of f^* as $p^*(Y|X)$, and then least squares can be viewed as discriminative learning, trying to find the true conditional distribution p^*. However, this way of learning p^* cannot be viewed as conditional likelihood maximization (note that there can be no notion of Gaussian noise here since Y is discrete, so the correspondence between maximum (conditoinal) likelihood and least squares is not applicable here. So using least squares is just a strange (and often not so good) way of trying to learn $p^*(y \mid x)$).

Naive Bayes and LDA fit into (and only into) Approach 1, Generative Models ...which should be clear enough.

Logistic Regression is an example of Approach 1 and Approach 3.

In Approach 1, the logistic regression model is the discriminative model corresponding to LDA, naive Bayes (and even some other) generative models. Note how different generative models can lead to the same discriminative model!

Under Approach 3, one can view logistic regression as a form of empirical risk minimization for the linear classification model (i.e. $\mathcal{F}$ is the set of all $f : \mathcal{X} \to \mathcal{Y}$ with linear classificationboundaries). The target is the 0/1-loss, but this is computationally hard to minimize; moroever, common algorithms for minimizing it are in a sense 'unstable'. So it is replaced during training by the conditional log-loss (which in this case can also be viewd as the 'logit loss'). That this is a good proxy for 0/1-loss is demonstrated by the fact that *if* the training set is 'linearly separable' (some linear classifier classifies the whole training set correctly), then the classifier minimizing the conditional log-loss in logistic regresion will also classify the whole training set correctly.

Perceptron, Support Vector Machine TO BE ANNOUNCED!