ML & Eco

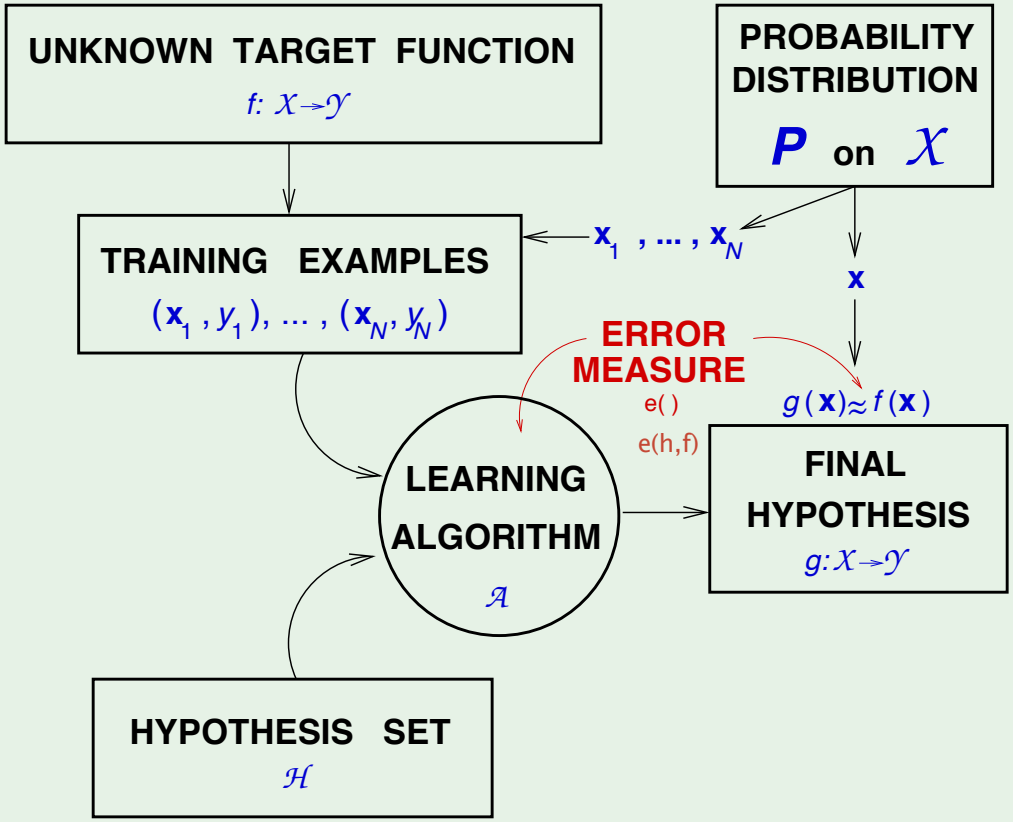
Econometric problems

- A Questions addressed by economists and social scientists
  - Can past information predict future stock returns at the firm level – large set of potentially correlated predictor variables,
  - Determining how an additional year of schooling affects earnings
  - Do matching grants increase charitable giving? Matching grants in effect decrease price.
  - Does exposure to advertising affect consumer choices?
  - What is the effect of demand of for a 1% increase in price of restaurant meals?
  - Do non-partisan phone calls encouraging people to vote increase voter turn out?
  - How does targeting pricing affect firm profits and consumer welfare?
  - How does teacher quality affect student test performance?

3 Themes

- Causation vs Curve fitting
  - causal effect
    - such as the effect of a training program,
    - a minimum wage increase,
    - or a price increase.
  - The researcher might check robustness of this parameter estimate by reporting two or three alternative specifications.
- Model Selection
  - Economics researcher picks a model and estimates it once
  - ML builds selection as part of algorithm
    - Cross Validation
- Valid confidence intervals for estimated effects

The learning diagram - with error measure

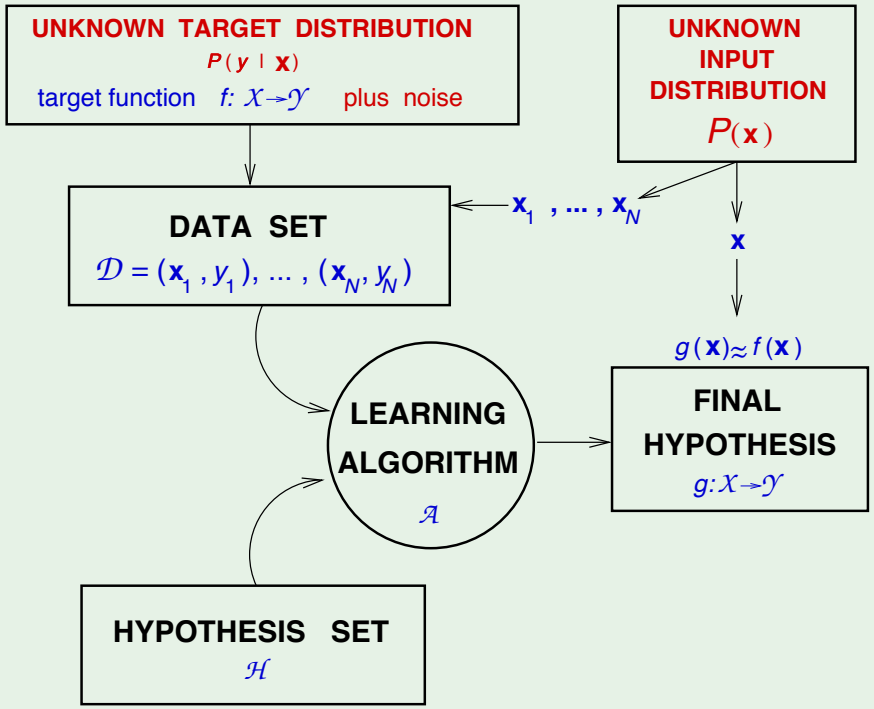


Probabilistic approach

Extend probabilistic role to all components

$P(\mathcal{D} \mid h = f)$ decides which h (likelihood)

How about $P(h = f \mid \mathcal{D})$?



An example

- An analyst wishes to estimate a model of consumer demand for different items,
  - common to model consumer preferences over characteristics of the items.
- Many items are associated with text descriptions as well as online reviews.
- Unsupervised learning could be used to discover items with similar product descriptions,
  - It could also be used to find subgroups of similar products.
- Unsupervised learning could further be used to categorize the reviews into types.
- An indicator for the review group could be used in subsequent analysis without the analyst having to use human judgement about the review content;
- Data would reveal whether a certain type of review was associated with higher consumer perceived quality, or not.
- An advantage of using unsupervised learning to create covariates is that the outcome data is not used at all; thus, concerns about spurious correlation between constructed covariates and the observed outcome are less problematic.

Another example

- Unsupervised learning can also be used to create outcome variables.
- examine the impact of Google's shutdown of Google News in Spain on the types of news consumers read.
- In this case, the share of news in different categories is an outcome of interest.
- Unsupervised learning can be used to categorize news in this type of analysis;
  - uses community detection techniques from network theory.

Another example

- Supervised machine learning typically entails using a set of features or covariates (X) to predict an outcome (Y).
- the goal is to construct $\hat{\mu}(x)$, on a training set, which is an estimator of $\mu(x) = E[Y|X = x]$, in order to do a good job predicting the true values of Y in an independent test dataset
- The observations are assumed to be independent, and the joint distribution of X and Y in the training set is the same as that in the test set.
- ML literature does not frame itself as solving estimation problems – so estimating $\mu(x)$ or $Pr(Y = k|X = x)$ is not the primary goal. Instead, the goal is to achieve goodness of fit in an independent test set by minimizing deviations between actual outcomes and predicted outcomes.
- In applied econometrics, we often wish to understand an object like $\mu(x)$ in order to perform exercises like evaluate the impact of changing one covariate while holding others constant.

Covariates

- Depending on the context, an independent variable is sometimes called a "predictor variable", regressor, covariate, "controlled variable", "manipulated variable", "explanatory variable", exposure variable (see reliability theory), "risk factor" (see medical statistics), "feature" (in machine learning and pattern recognition) or "input variable."[11][12] In econometrics, the term "control variable" is usually used instead of “covariate”.

- Depending on the context, a dependent variable is sometimes called a "response variable", "regressand", "criterion", "predicted variable", "measured variable", "explained variable", "experimental variable", "responding variable", "outcome variable", "output variable" or "label".[12]

Ml vs Econometrics methods

- one common feature of many ML methods is that they use data-driven model selection.
  - Analyst provides the list of covariates or features, but the functional form is at least in part determined as a function of the data
- rather than performing a single estimation (as is done, at least in theory, in econometrics),
- so that the method is better described as an algorithm that might estimate many alternative models and then select among them to maximize a criterion.
- The ML literature uses a variety of techniques to balance expressiveness against over-fitting. The most common approach is cross-validation

Econometrics

- Researcher specifies one model, estimates the model on the full dataset, and relies on statistical theory to estimate confidence intervals for estimated parameters.
- The focus is on the estimated effects rather than the goodness of fit of the model.
- Researchers often check dozens or even hundreds of alternative specifications behind the scenes, but rarely report this practice because it would invalidate the confidence intervals reported (due to concerns about multi-ple testing and searching for specifications with the desired results).

Causal Inference

- Instrumental variables are used by economists when they wish to learn a causal effect, for example the effect of a price on a firm's sales, but they only have access to observational (non-experimental) data.
- An instrument in this case might be an input cost for the firm that shifts over time, and is unrelated to factors that shift consumer's demand for the product (such demand shifters can be referred to as "con-founders" becaues the affect both the optimal price set by the firm and the sales of the product).
- It is very common to see that a predictive model (e.g. least squares regression) might have very high explanatory power (e.g. high R^2), while the causal model (e.g. instrumental variables regression) might have very low explanatory power (in terms of predicting outcomes). In other words, economists typically abandon the goal of accurate prediction of outcomes in pursuit of an unbiased estimate of a causal parameter of interest.