Statistical Machine Learning Introduction

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Standard statistical notation

	Variables				
Cases	x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
	x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_{N1}	x_{N2}	$\cdots$	x_{Np}	y_N

$$\underset{N\times p}{\mathbf{X}} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_N' \end{pmatrix}, \underset{N\times 1}{\mathbf{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \text{ and } \boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Y})$$

Objective and “classical statistics” vs “machine learning” perspectives

- Identification, description, and enabling the use of simple/low-dimensional/low-order structure in the data array

- Data are scarce vs data are plentiful

- Quantification of what is known about (probability) models used is central vs no real interest in this issue

Types of statistical machine learning problems

- Supervised learning/prediction

$$\hat{y} = \hat{f}\left(\mathbf{x}\right)$$

  - continuous target y or $y \in \{1, 2, \ldots, K\}$ for classification/pattern recognition

- Unsupervised learning

  - detailing relationships between the entries in $\mathbf{x}$ or commonalities among sets of cases

What is really new here (particularly in prediction)?

- "Big" datasets allow the creation of complex/flexible prediction methods
- With large p, datasets are inevitably sparse and the possible complexity of predictors explodes … this is "the curse of dimensionality"
- The consequent possibility of "overfit" requires that predictor complexity must be matched to the real information content in a training set
  - the effectiveness of a prediction methodology can only be reliably judged in terms of performance on a "holdout" sample ("training" and "testing" sets cannot be the same)

Reduction of what is known to an $N \times (p+1)$ training set for prediction

- This is typically highly labor-intensive (often 80% of person hours in corporate projects?)
  - assembling case information from many sources
  - data cleaning
  - data formatting

- This governs/limits what can be done in prediction

- Technically speaking, all that follows treats the training and test cases (the "pairs" $(\mathbf{x}_i, y_i)$) as iid/random draws from a fixed universe

Standardization of quantitative features

- In general, "interval level"/quantitative features are much more easily used in prediction than are ordinal or categorical ones

- These often naturally come with corresponding units
  - sometimes this is no logical problem
  - often, however, different units for different features creates logical problems

- A way to avoid inconsistencies is to standardize coordinates of $\mathbf{x}$

$$x' \equiv \frac{x - \bar{x}}{s_x}$$