Naivity can be good:
a theoretical study of naive regression

C. Andy Tsao* and Li-Yin Chen

Institute of Statistics/Department of Applied Math
National Dong Hwa University

August, 2012
Harbin, China

Outline

Supervised learning

- Training data:
  $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X} = \mathcal{R}^p$ and $y_i \in \mathcal{Y} = \{\pm 1\}$
- Testing (generalization) data: $\{(x'_j, y'_j)\}_{j=1}^{m}$
- Distribution: $(x_i, y_i) \overset{from}{\leftarrow} (X, Y) \overset{iid}{\sim} P_{X,Y}$
- Machine or classifier: Find $G \in \mathcal{F}$ such that $\widehat{G} : \mathcal{X} \to \mathcal{Y}$

Supervised learning

- Training data:
  $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X} = \mathcal{R}^p$ and $y_i \in \mathcal{Y} = \{\pm 1\}$
- Testing (generalization) data: $\{(x'_j, y'_j)\}_{j=1}^m$
- Distribution: $(x_i, y_i) \overset{from}{\leftarrow} (X, Y) \overset{iid}{\sim} P_{X,Y}$
- Machine or classifier: Find $G \in \mathcal{F}$ such that $\widehat{G} : \mathcal{X} \to \mathcal{Y}$
- Training error:

$$TE = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{[y_i \neq \widehat{G}(x_i)]} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{[y_i \widehat{G}(x_i) < 0]}$$

- Testing (generalization) error:

$$\widehat{GE} = \frac{1}{m}\sum_{j=1}^m \mathbf{1}_{[y'_j \widehat{G}(x'_j) < 0]} \text{ and } GE = E_{X,Y}\{\mathbf{1}_{[YG(X)<0]}\}$$

Regression

- Data:
  $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \mathcal{R} = (-\infty, \infty)$
- Distribution: $(x_i, y_i) \overset{from}{\longleftarrow} (Y_i|x_i) \overset{indep.dist}{\sim} P_{Y|x}$
- Machine or Regression: Find $G \in \mathcal{F}$ such that $\widehat{G} : \mathcal{X} \rightarrow \mathcal{Y}$

Regression

- Data:
  $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \mathcal{R} = (-\infty, \infty)$
- Distribution: $(x_i, y_i) \overset{from}{\leftarrow} (Y_i | x_i) \overset{indep.dist}{\sim} P_{Y|x}$
- Machine or Regression: Find $G \in \mathcal{F}$ such that $\widehat{G} : \mathcal{X} \to \mathcal{Y}$
- Sum of Square Errors, $G(X) = \beta_0 + \beta' X$,
  $\beta = (\beta_1, \cdots, \beta_p)', X = (X_1, X_2, \cdots, X_p)'$.

$$SSE = ||Y - \hat{Y}||^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{G}(x_i))^2$$

where

$$\widehat{G}(x) = \hat{\beta}_0 + \hat{\beta}' x$$

- LSE: $\hat{\beta}_0, \hat{\beta}$; $\hat{Y}$ projection of Y onto CS of the design matrix.

Naive regression

- Data:
  $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \{-1, 1\}$
- Distribution: $(x_i, y_i) \overset{from}{\leftarrow} (Y_i|x_i) \overset{indep.dist}{\sim} P_{Y|x}$
- Machine or Regression: Find $G \in \mathcal{F}$ such that $\widehat{G} : \mathcal{X} \to \mathcal{Y}$

Naive regression

- Data:
  $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \{-1, 1\}$
- Distribution: $(x_i, y_i) \overset{from}{\leftarrow} (Y_i|x_i) \overset{indep.dist}{\sim} P_{Y|x}$
- Machine or Regression: Find $G \in \mathcal{F}$ such that $\widehat{G} : \mathcal{X} \to \mathcal{Y}$

$$\widehat{G}(x) = \hat{\beta}_0 + \hat{\beta}' x$$

- TE and GE
- LSE: $\hat{\beta}_0, \hat{\beta}$; $\hat{Y}$ projection of Y onto CS of the design matrix.

Naivity of naive regression

- *Labels/factors* in $\mathcal{Y}$ as *numbers*
- G is linear
- LSE $\hat{\beta}_0, \hat{\beta}$

The good

- Straightforward/Naive: Similar to regression/general linear model
- Easy implementation: say `glm` in `R`
- Adoption of tricks or methods in regression: variable selection

The bad

- Violation of Gauss-Markov Theorem
- LSE $\nrightarrow$ Errors in classification
- $\hat{y} \notin [0, 1]$

The ugly/beautiful-1

Proposition (NR-LDA equivalence)

Let $\mathcal{Y} = \{\frac{-n_1}{n}, \frac{n_2}{n}\}$, $\hat{\beta}_0, \hat{\beta}$ *minimizes* $\sum_{i=1}^{n}(y_i - \beta_0 - \beta' x_i)^2$ *and* $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}' x$. *Then*

1. $\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$
2. *If the data is completely balanced, i.e.* $n_1 = n_2$, *then* $\hat{f}(x) > 0$ *iff LDA classifies the case to class* 2.

Recall LDA classifies the case to class 2 if

$$x'\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 - \hat{\mu}_1)'\hat{\Sigma}^{-1}(\hat{\mu}_2 + \hat{\mu}_1) + \log(\hat{\pi}_1) - \log(\hat{\pi}_2),$$

and class 1 otherwise with $\hat{\pi}_i = n_i/n, i = 1, 2$.
Ripley (1996), Fisher (1936),
Hastie, Tibishirani and Friedman (2009).

The ugly/beautiful-2

- $\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$ if $\mathcal{Y} = \{-1, 1\}$ or any distinct coding of two classes, e.g. $\{0, 1\}$ or $\{1, 2\}$.
- The decision hyperplanes of NR and LDA share the same normal vector (subject to normalization)
- For completely balanced data,
  i.e. $n_1 = n_2$, NR is equivalent to LDA.

The ugly/beautiful-3

Proposition (Class estimates)

For $k = 1, \cdots, K$, let t_k be an indicator vector with the k-th entry equals 1 and all other entries equal zero. Let $\mathbf{y} = (y_1, \cdots, y_K)'$ then

$$argmin_k ||t_k - \mathbf{y}|| = argmax_k \ y_k (= k_0).$$

$$||t_k - \mathbf{y}||^2 - ||t_{k_0} - \mathbf{y}||^2 = 2(t_{k_0} - t_k)'\mathbf{y} = 2(y_{k_0} - y_k) \geq 0. \quad (1)$$

Class estimates: $\sum_k y_k = 1, y_k \in [0, 1]$.

Hastie, Tibishirani and Friedman (2009).

The ugly/beautiful-4

- No assumption on $\sum_k y_k = 1$ nor $y_k \in [0,1]$ is needed
- Interpret $\mathbf{y} = (y_1, \cdots, y_K)' = \hat{f}(x)$ as class probabilities
- For $\mathcal{Y} = \{0,1\}$, NR $\sim$ regression on the indicator response matrix, e.g.

Let $Y^c = 1 - Y$ $\quad (Y, \quad Y^c | X_1, X_2)$

$$\begin{pmatrix} 1 & 0 & | & 2 & 3 \\ 0 & 1 & | & 1 & 5 \\ 1 & 0 & | & 3 & 2 \end{pmatrix}$$

Implications

- NR is invariant wrt different codings of Y, e.g. $\mathcal{Y} = \{0,1\}, \{-1,1\}$ or $\{1,2\}$.
- NR and LDA have the same ROC curve (hyperplanes share the same normal vector)
- If $n_1 = n_2$, NR=LDA
- OK even if $\hat{y} \notin [0,1]$
- LDA with some categorical X's: similar to those in GLM
- LSE: $\hat{Y} = \hat{G}(X) = \hat{\beta}_0 + \hat{\beta}'X$ is unique but $(\hat{\beta}_0, \hat{\beta})$ is not.
- Variable selection
- Implementation: Least Angle Regression (`LARS` package, `R`)

Efron, et al (2004).

Conclusion and Discussion

- NR is an easy classifier and relates closely to LDA and indicator matrix regression
- Alternative implementation for LDA for binary classification with nearly balanced data
- Tricks and ideas in GLM can be readily adapted
- Kernel FDA, extension to multi-class classification
- LDA vs. NR: Plug-in Population Bayes Rule vs. Sample version decision rule.

Thanks for your attention!

- EFRON, B., HASTIE, T., JOHNSTONE, I. AND TIBISHIRANI, R. (2004).
  Least angle regression. Annals of Statistics **32**, 407–499.
- FISHER, R.A. (1936).
  The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.
- HASTIE, T., TIBISHIRANI, R. AND FRIEDMAN, J. (2009).
  The elements of statistical learning. 2nd Edition, Springer.
- MIKA, S., RÄTSCH, G., WESTON, J., SCHÖLKOPF, B. AND MÜLLER, K.R. (2012).
  Fisher Discriminant analysis with kernels. IEEE Transactions on Neural Networks, **100**, 1000–1017.
- RIPLEY, B.D. (1996).
  Pattern recognition and neural networks.
  Cambridge University Press.