Multi-View Regression via Canonincal Correlation Analysis

Dean P. Foster
U. Pennsylvania

(Sham M. Kakade of TTI)

Model and background

$$(\forall i \leq n) \qquad y_i = \sum_{i=j}^{p} X_{ij}\beta_j + \epsilon_i \quad \epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$$

Data mining and Machine learning: $p \gg n$

Model and background

$$(\forall i \leq n) \qquad y_i = \sum_{i=j}^{p} X_{ij}\beta_j + \epsilon_i \quad \epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$$

Can't fit model if $p \gg n$:

- Trick: assume most β_i are in fact zero
- Variable selection:

$$\hat{\beta}_i^{\text{RIC}} = \begin{cases} 0 & \text{if } |\hat{\beta}_i| \leq SE_i\sqrt{2\log p} \\ \hat{\beta}_i & \text{otherwise} \end{cases}$$

- Basically just stepwise regression and Bonferroni
  - Can be justified by "risk ratios" (Donoho and Johnstone '94, Foster and George '94)

Model and background

$$(\forall i \leq n) \qquad y_i = \sum_{i=j}^{p} X_{ij}\beta_j + \epsilon_i \quad \epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$$

I've played with lots of alternatives:

- FDR instead of RIC:
  - $\sqrt{2 \log p} \to \sqrt{2 \log(p/q)}$
  - empirical Bayes (George and Foster, 2000)
  - Cauchy prior (Foster and Stine, 200x)
- regression $\to$ logistic regression
- IID $\to$ independence
- independence $\to$ block independence (with Dongyu Lin)

Model and background

$$(\forall i \leq n) \qquad y_i = \sum_{i=j}^{p} X_{ij}\beta_j + \epsilon_i \quad \epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$$

Where do this many variables come from?

- Missing value codes
- Interactions
- Transformations
- Example (with Bob)
  - Personal Bankruptcy
  - 350 basic variables
  - all interactions, missing value codes, etc lead to 67,000 variables
  - about 1 million clustered cases
  - Ran stepwise logistic regression using FDR

Model and background

$$(\forall i \leq n) \qquad y_i = \sum_{i=j}^{p} X_{ij}\beta_j + \epsilon_i \quad \epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$$

Summary of current state of the art:

- We can generate many non-linear X's
- We can select the good ones large lists
- Isn't the problem "solved"?

Model and background

$$(\forall i \leq n) \qquad y_i = \sum_{i=j}^{p} X_{ij}\beta_j + \epsilon_i \quad \epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$$

There is always room for finding new X's

New features

- Current methods of finding X's are non-linear
- Can we find "new" linear combinations of existing X's?
  - Hope, use linear theory
  - Hope, fast CPU
  - Hope, new theory

Semi-supervised

Semi-supervised learning is:

- Y's are expensive
- X's are cheap
- We get n rows of Y
- But also m free rows of just X's
- Called, semi-supervised learning
- Can this help?

Usual data table for data mining

$$\begin{bmatrix} Y \\ \\ (n \times 1) \\ \\ \end{bmatrix} \begin{bmatrix} X \\ \\ (n \times p) \\ \\ \end{bmatrix}$$

with $p \gg n$

With unlabeled data

m rows of unlabeled data:

$$\begin{bmatrix} Y \\ \\ n \times 1 \\ \\ \end{bmatrix} \quad \begin{bmatrix} X \\ \\ (n+m) \times p \\ \\ \\ \\ \\ \end{bmatrix}$$

With alternative X's

m rows of unlabeled data, and two sets of equally useful X's:

$$\begin{bmatrix} Y \\ n \times 1 \\ \end{bmatrix} \quad \begin{bmatrix} X \\ (n+m) \times p \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} Z \\ (n+m) \times p \\ \\ \\ \end{bmatrix}$$

With: $m \gg n$

Examples

- Person identification
  - Y = identity
  - X = Profile photo
  - Z = front photo
- Topic identification (medline)
  - Y = topic
  - X = abstract
  - Z = text
- The web:
  - Y = classification
  - X = content (i.e. words)
  - Z = hyper-links
- We will call these the multi-view setup

A Multi-View Assumption

Define

$$\begin{aligned}\sigma_x^2 &= E[Y - E(Y|X)]^2 \\ \sigma_z^2 &= E[Y - E(Y|Z)]^2 \\ \sigma_{x,z}^2 &= E[Y - E(Y|X,Z)]^2\end{aligned}$$

(We will take conditional expectations to be linear)

Assumption

Y,X, and Z satisfy the α-multiview assumption if:

$$\begin{aligned}\sigma_x^2 &\leq \sigma_{x,z}^2(1+\alpha) \\ \sigma_z^2 &\leq \sigma_{x,z}^2(1+\alpha)\end{aligned}$$

- In other words, $\sigma_x^2 \approx \sigma_z^2 \approx \sigma_{x,z}^2$
- Views X and Z are redundant (i.e. highly collinear)

The Multi-View Assumption in the Linear Case

- The views are redundant.
- Satisfied if each view predict Y well.
- No conditional independence assumptions (i.e. Bayes nets)
- No coordinates, norm, eigenvalues, or dimensionality assumptions.

Both estimators are similar

Lemma

Under the α-multiview assumption

$$E[(E(Y|X) - E(Y|Z))^2] \leq 2\alpha\sigma^2$$

- Idea: find directions in X and Z that are highly correlated
- CCA solves this problem already!

What if we run CCA on X and Z?

CCA = canonical correlation analysis

- Find the directions that are most highly correlated
- Very close to PCA (principal components analysis)
- Generates coordinates for data
- End up with canonical coordinates for both X's and Z's
- Numerically an Eigen-value problem

CCA form

Running CCA on X and Z generates a coordinate space for both of them. We will call this "CCA form."

CCA form

Running CCA on X and Z generates a coordinate space for both of them. We will call this "CCA form."

Definition

X_i, and Z_j, are in CCA form if

- X_i are orthonormal
- Z_i are orthonormal
- $X_i^T Z_j = 0$ for $i \neq j$
- $X_i^T Z_i = \lambda_i, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

(This is the output of running CCA on the original X's and Z's.)

CCA form as a covariance matrix

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{XX} & \Sigma_{XZ} \\ \hline \Sigma_{ZX} & \Sigma_{ZZ} \end{array}\right] \rightarrow \left[\begin{array}{c|c} \mathrm{I} & D \\ \hline D & \mathrm{I} \end{array}\right]$$

The *canonical correlations* are λ_i:

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots \\ 0 & \lambda_2 & 0 & \dots \\ 0 & 0 & \lambda_3 & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

The Main Result

Theorem

Let $\hat{\beta}$ be the Ridge regression estimator with weights induced by the CCA. Then

$$Risk(\hat{\beta}) \leq \left(5\alpha + \frac{\sum \lambda_i^2}{n}\right)\sigma^2$$

The Main Result

Theorem

Let $\hat{\beta}$ be the Ridge regression estimator with weights induced by the CCA. Then

$$Risk(\hat{\beta}) \leq \left(5\alpha + \frac{\sum \lambda_i^2}{n}\right)\sigma^2$$

CCA-ridge regression is to minimize least squares plus a penalty of:

$$\sum_i \frac{1-\lambda_i}{\lambda_i}\beta_i^2$$

- Large penalties in the less correlated directions.
- λ_i's are the correlations
- A shrinkage estimator.

The Main Result

Theorem

Let $\hat{\beta}$ be the Ridge regression estimator with weights induced by the CCA. Then

$$Risk(\hat{\beta}) \leq \left(5\alpha + \frac{\sum \lambda_i^2}{n}\right)\sigma^2$$

Recall α is the multiview property:

$$\begin{aligned} \sigma_x^2 &\leq \sigma_{x,z}^2(1+\alpha) \\ \sigma_z^2 &\leq \sigma_{x,z}^2(1+\alpha) \end{aligned}$$

The Main Result

Theorem

Let $\hat{\beta}$ be the Ridge regression estimator with weights induced by the CCA. Then

$$Risk(\hat{\beta}) \leq \left(5\alpha + \frac{\sum \lambda_i^2}{n}\right)\sigma^2$$

- 5α is the bias
- $\frac{\sum \lambda_i^2}{n}$ is variance

Doesn't fit my style

- I like feature selection!
- On to theorem 2

Alternative version

Theorem

For $\hat{\beta}$ be the CCA-testimator:

$$Risk(\hat{\beta}) \leq \left(2\sqrt{\alpha} + \frac{d}{n}\right)\sigma^2$$

where d is the number of λ_i for which $\lambda_i \geq 1 - \sqrt{\alpha}$.

Alternative version

Theorem

For $\hat{\beta}$ be the CCA-testimator:

$$Risk(\hat{\beta}) \leq \left(2\sqrt{\alpha} + \frac{d}{n}\right)\sigma^2$$

where d is the number of λ_i for which $\lambda_i \geq 1 - \sqrt{\alpha}$.

The CCA testimator:

$$\widehat{\beta}_i = \begin{cases} \mathrm{MLE}(\beta_i) & \text{if } \lambda_i \geq 1 - \sqrt{\alpha} \\ 0 & \text{else} \end{cases} \tag{1}$$

Alternative version

Theorem

For $\hat{\beta}$ be the CCA-testimator:

$$Risk(\hat{\beta}) \leq \left(2\sqrt{\alpha} + \frac{d}{n}\right)\sigma^2$$

where d is the number of λ_i for which $\lambda_i \geq 1 - \sqrt{\alpha}$.

Do we need to know α?

- We can try features in order
- Use promiscuous rule to add variables (i.e. AIC)
- Will do as well as theorem, and possibly much better
- Doesn't mix all that well with stepwise regression

Conclusions

- Trade off between two theorems?
- Experimental work?

Conclusions

- Trade off between two theorems?
- Experimental work? Soon!