
Learning gradients: predictive models that infer geometry and dependence

Q. Wu, J. Guinney, M. Maggioni, S. Mukherjee
Preprint, JMLR, 2007

Discussion Notes prepared by

Xuejun Liao

18 October 2007

Background and motivation

- Dimensionality reduction, manifold learning
- Are these relevant to prediction?
- How to take target variable into consideration?
- Learning prediction model and manifold simultaneously

Basic ideas (1/2)

- Discriminative models (regression/classification): $p(Y|X)$
- Generative models (inverse regression): $p(X|Y)$
- Learning gradients – jointly estimate the **regression function**

$$f_r(x) = \mathbb{E}_Y [Y|X = x]$$

and the **covariation of the inverse regression**

$$\Omega_{X|Y} = \text{cov}(\mathbb{E}(X|Y))$$

- Gradient outer product (GOP) matrix:

$$\Gamma = \mathbb{E} \left\{ [\nabla_x f_r] [\nabla_x f_r]^T \right\}$$

Basic ideas (2/2)

Proposition 1. Assume $y = \beta^T x + \epsilon$, $\epsilon \sim N(0, \sigma_\epsilon^2)$. Given $\Omega_{X|Y} = \text{cov}(\mathbb{E}(X|Y))$, $\sigma_Y^2 = \text{var}(Y)$, $\Sigma_X = \text{cov}(X)$, the GOP matrix is

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1}$$

Corollary 2. Given $\mathcal{X} = \cup_{i=1}^{\mathcal{I}} R_i$, $f_r(x) = \beta_i^T x + \epsilon_i \forall x \in R_i$, $\mathbb{E}\epsilon_i = 0$, $\Omega_i = \text{cov}(\mathbb{E}(X \in R_i|Y))$, $\sigma_Y^2 = \text{var}(Y|X \in R_i)$, $\Sigma_i = \text{cov}(X \in R_i)$, the GOP matrix can be defined in terms of the local quantities

$$\Gamma = \sum_{i=1}^{\mathcal{I}} \rho_X(R_i) \sigma_i^2 \left(1 - \frac{\sigma_{\epsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1} \Omega_i \Sigma_i^{-1}$$

where $\rho_X(R_i) = \int_{x \in R_i} \rho_X(x) dx$

Estimating gradients (1/2)

- Taylor expansion of the regression function

$$f_r(u) \approx f_r(x) + \nabla f_r(x) \cdot (u - x), \text{ for } u \approx x$$

- When evaluated at data samples,

$$f_r(x_i) \approx f_r(x_j) + \nabla f_r(x_j) \cdot (x_i - x_j), \text{ for } x_i \approx x_j$$

- In regression, given data $D = \{(y_i, x_i)\}_{i=1}^n$, simultaneously estimate

$$(\vec{f}_D, f_D) = \arg \min_{(\vec{f}, f) \in \mathcal{H}_K^{p+1}} \left(\varepsilon_D(\vec{f}, f) + \lambda_1 \|f\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right)$$

where, $\varepsilon_D(\vec{f}, f) = \frac{1}{n^2} \sum_{i,j=1}^n w_{ij}^s \left(y_i - \left[f(x_j) + \vec{f}(x_j) \cdot (x_i - x_j) \right] \right)^2$,
 $\|\cdot\|_K$ is the norm in reproducing kernel Hilbert space (RKHS),

$$w_{ij}^s = \exp(-\|x_i - x_j\|^2 / 2s^2)$$

Estimating gradients (2/2)

- In classification, given data $D = \{(y_i, x_i)\}_{i=1}^n$, simultaneously estimate

$$(\vec{f}_D, f_D) = \arg \min_{(\vec{f}, f) \in \mathcal{H}_K^{p+1}} \left(\varepsilon_D^\phi(\vec{f}, f) + \lambda_1 \|f\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right)$$

where, $\varepsilon_D^\phi(\vec{f}, f) = \frac{1}{n^2} \sum_{i,j=1}^n w_{ij}^s \phi \left(y_i \left[f(x_j) + \vec{f}(x_j) \cdot (x_i - x_j) \right] \right)$,
 $\phi(t) = \ln(1 + e^{-t})$ is the logistic loss function.

- By the representer theorem (Wahba, 1990), the solutions in both cases are, for $\alpha_{i,D} \in \mathbb{R}$ and $c_{i,D} \in \mathbb{R}^p$,

$$f_D(x) = \sum_{i=1}^n \alpha_{i,D} K(x, x_i), \quad \vec{f}_D(x) = \sum_{i=1}^n c_{i,D} K(x, x_i)$$

- The GOP matrix is estimated as $\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^n \vec{f}_D(x_i) \otimes \vec{f}_D(x_i)$

Joint features reduction & prediction (1/2)

- Feature selection:

$$\text{ranking of the } i\text{-th feature} = \hat{\Gamma}_{ii} \approx \mathbb{E}_X \left(\frac{\partial f_r}{\partial x^i} \right)^2$$

- Linear feature construction — eigenvalue decomposition of GOP

$$\hat{\Gamma} = \sum_{i=1}^n \lambda_i u_i u_i^T, \quad \lambda_1 \geq \dots \geq \lambda_m \geq \lambda_{m+1} \approx \dots \approx \lambda_p \approx 0$$

then u_1, \dots, u_m define the $m \ll n$ new features most relevant to the prediction

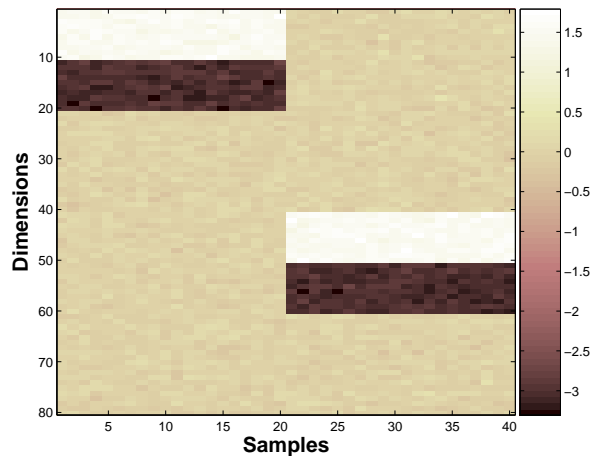
Joint features reduction & prediction (2/2)

- Nonlinear dimension reduction — gradient based diffusion maps
 - Function adapted graph affinity matrix

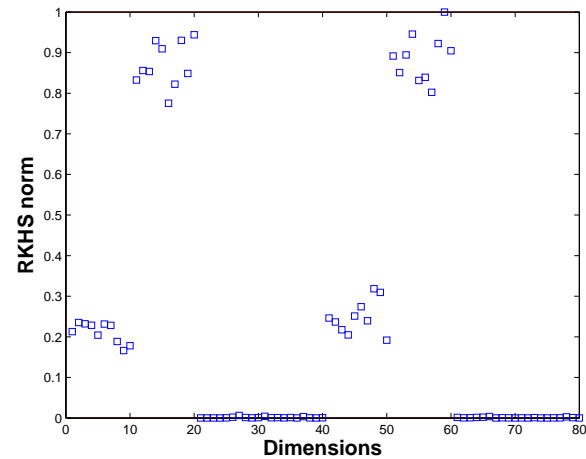
$$W_f(x_i, x_j) = \exp \left(-\frac{\|x_i - x_j\|^2}{\sigma_1} - \frac{\|f(x_i) - f(x_j)\|^2}{\sigma_2} \right)$$

constructed for labeled as well as unlabeled data

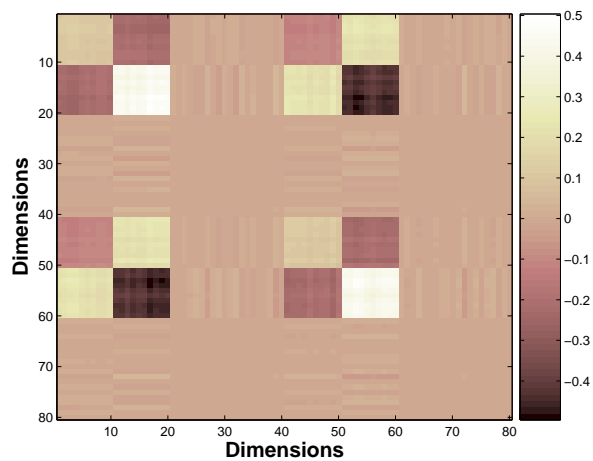
- $f(x_i) - f(x_j) \approx \vec{f}_D(x_i)(x_j - x_i)$ or
 $f(x_i) - f(x_j) \approx \frac{1}{2}[\vec{f}_D(x_i)(x_j - x_i) + \vec{f}_D(x_j)(x_i - x_j)],$
with \vec{f}_D estimated from labeled data
- Construct diffusion maps based on W_f



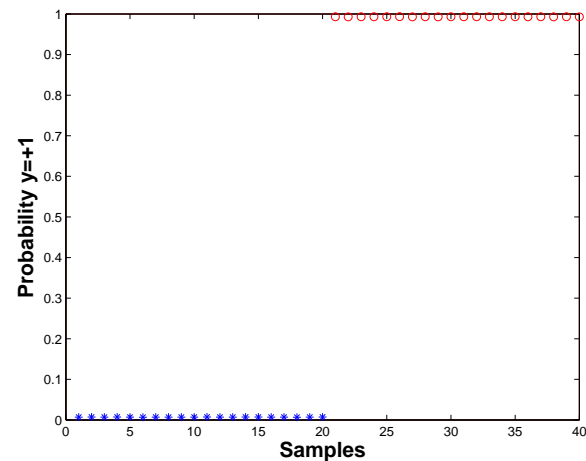
(a)



(b)

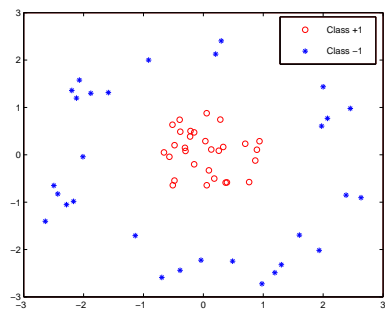


(c)

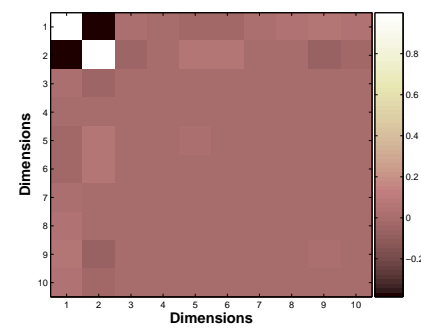


(d)

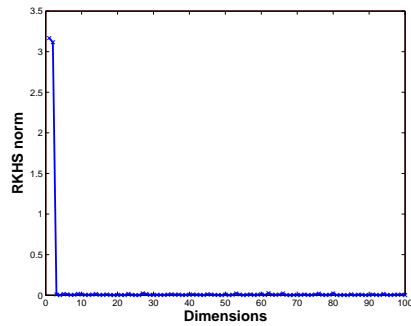
Figure 1: a) The data matrix \mathbf{x} where each sample corresponds to a column and the first twenty samples correspond to class -1 and the second twenty to class $+1$, b) the RKHS norm for each dimension, c) the empirical covariance matrix, d) the predicted class probabilities on the training data.



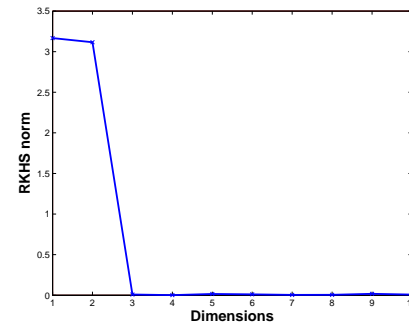
original dimensionality = 200 (a)



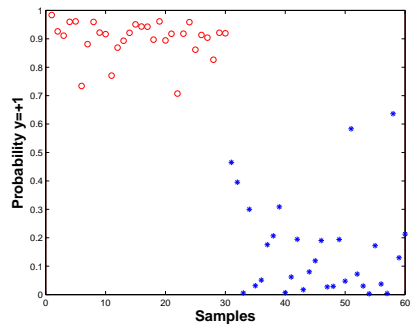
(b)



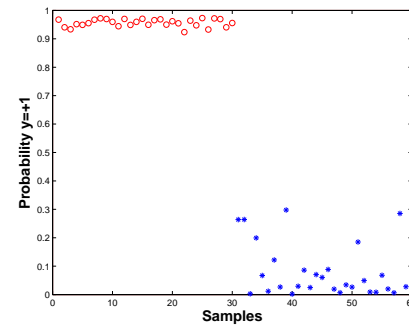
(c)



(d)



(e)



(f)

Figure 3: a) The first two dimensions of the data matrix class +1 are the circles and class -1 are the stars, b) the empirical covariance matrix for the first 10 dimensions, c) the RKHS norm for the first 100 dimensions, d) the RKHS norm for the first 10 dimensions, e) the predicted class probabilities on the training data with no feature selection again circles are class +1 and stars are class -1, f) the predicted class probabilities on the training data with feature selection.