Introduction to Kernel Methods I

Partha Niyogi

Mikhail Belkin

The University of Chicago

Interdisciplinary Subject

Variously called

- Pattern Recognition
- Machine Learning
- Classification/Regression

Many disciplines

- Engineering
- Computer Science
- Statistics
- Mathematics

Learning from examples

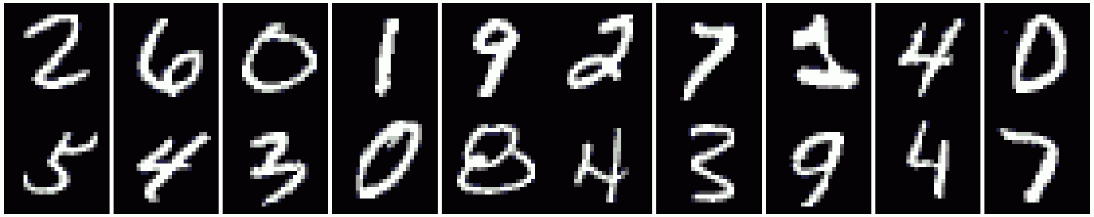
$$X = \begin{cases} \text{Pattern Space} \\ \text{Instance Space} \\ \text{Example Space} \end{cases} \qquad \mathbb{R}^n, \mathcal{M}, \{-1,+1\}^n, \Sigma^*$$

$$Y = \begin{cases} \text{Label Space} \\ \text{Prediction Space} \\ \text{Response Space} \end{cases} \qquad \mathbb{R}^n, \{-1,+1\}, \{1,\ldots,n\}$$

Examples (x, y)

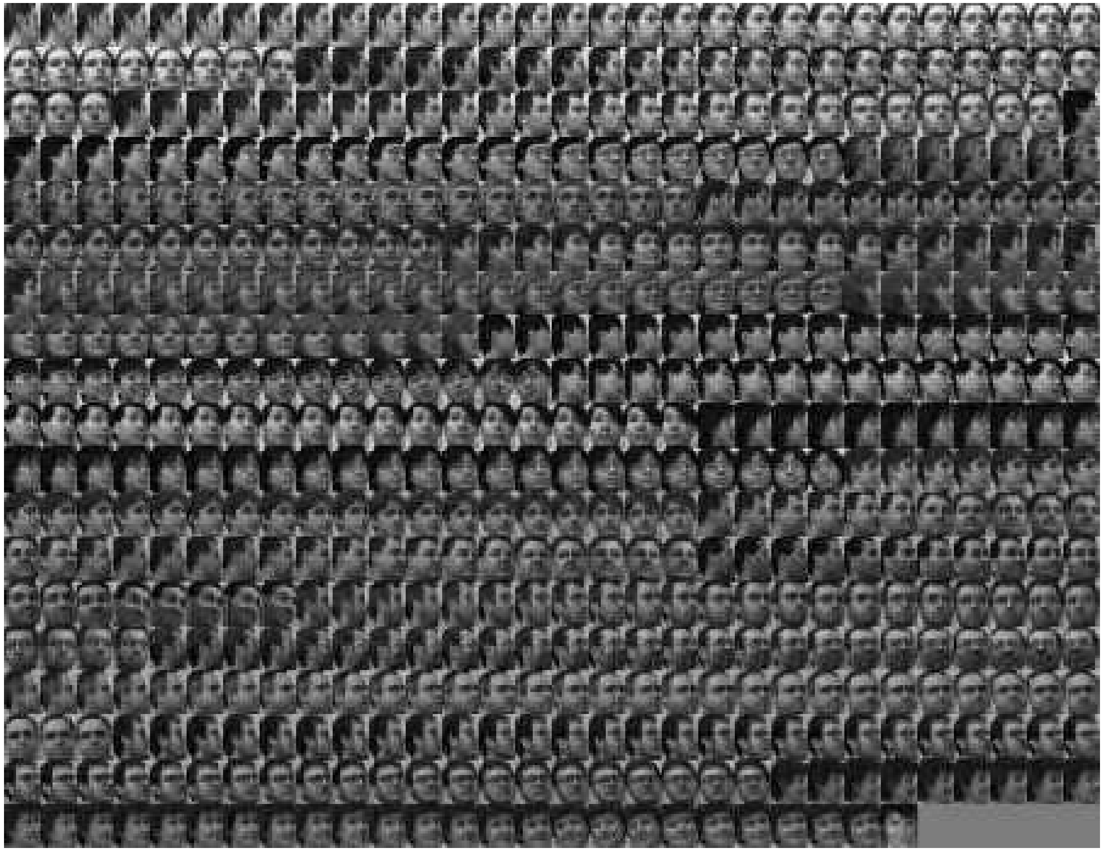
A Digit Recognition Example

Examples:



Predict class of new data point.

A Face Recognition Example



Predict class of new data point.

Finance

LUCENT TECHNOLOGY
as of 13-May-2005
Splits:
4.5
4.0
3.5
3.0
2.5
2.0
Jul04
Sep04
Nov04
Jan05
Mar05
May05
200
150
100
50
0
Millions
Copyright 2005 Yahoo! Inc.
http://finance.yahoo.com/

$$X = \Sigma^*$$

He ran from there with his money.

He his money with from there ran.

Learn $g : \Sigma^* \to \{-1, 1\}$.

Pattern Recognition

P on $X \times Y$ $\quad\quad$ $X = \mathbb{R}^N \quad Y = \{-1, 1\}$ or $\mathbb{R}$

(x_i, y_i) labeled examples

find $f : X \to Y$ $\quad\quad$ *Ill Posed*

$Y = \{-1, +1\}$: Misclassification Loss

$$V(f(x), y) = \begin{cases} 1 & \textbf{if } f(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{E}\left[V(f(x), y)\right] = \Pr[f(x) \neq y] = \text{ Average Error}$$

Bayes Optimal

Suppose P is known to you.
Suppose all measurable functions are available.

$$\min_f \mathbf{E}\left[V(f(x), y)\right] = \Pr[f(x) \neq y]$$

$$f_*(x) = \begin{cases} +1 & \Pr[y = +1 \mid x] \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

is the minimizer and Bayes optimal classifier.

$$V(f(x), y) = (y - f(x))^2$$

Minimizer of least squares is the regression function

$$f_*(x) = \mathbf{E}\left[y \mid x\right] = \int_Y y P(y|x) dy$$

sign $(f_*(x)) =$ Bayes Optimal Classifier

Empirical Risk Minimization

Choose a class $\mathcal{F}$ of functions $X \mapsto Y$

Solve

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i)$$

Ordinary Least Squares

$$\mathcal{F} = \{\mathbf{w} \cdot \mathbf{x} \mid \mathbf{w} \in \mathbb{R}^k\}$$

$$\min_{\mathbf{w}} \sum_i (y_i - (\mathbf{w} \cdot \mathbf{x}_i))^2$$

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

Differentiating with respect to $\mathbf{w}$ and setting to 0,

$$\mathbf{X}^T\mathbf{X}\mathbf{w}_* = \mathbf{X}^T y$$

Minimum Norm Solution

What if $\mathbf{X}^T\mathbf{X}$ is not full rank, i.e., not invertible?

$$H = \{\mathbf{w} \mid \mathbf{w} \text{ is minimizer}\}$$

Pick

$$\mathbf{w}_* = \min_{\mathbf{w} \in H} \mathbf{w} \cdot \mathbf{w}$$

Regularized Least Squares

$$\mathbf{w}_* = \arg\min_{\mathbf{w}} \frac{1}{n} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 + \gamma \mathbf{w} \cdot \mathbf{w}$$

$$w_* = [\mathbf{X}^T \mathbf{X} + \gamma I]^{-1} \mathbf{X}^T \mathbf{y}$$

Ridge Regression

$$\sum_{i=1}^{n} V(\mathbf{w}\cdot\mathbf{x}_i + b, y_i) + \gamma \mathbf{w}\cdot\mathbf{w}$$

where V is the hinge loss given by

$$V(f(x), y) = \begin{cases} 0 & \text{if } \; yf(x) \geq 1 \\ (1 - yf(x)) & \text{otherwise} \end{cases}$$

$$\underset{\{\mathbf{w},\xi_i\}}{\operatorname{argmin}}\, \mathbf{w}\cdot\mathbf{w} + \gamma \sum_{i=1}^{n} \xi_i$$

subject to

$$y_i\left(\mathbf{w}\cdot\mathbf{x}_i + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

One can show

$$\mathbf{w}^* = \sum \alpha_i x_i$$

Beyond linear functions

We would like a richer class $\mathcal{F}$ of functions with which to make predictions.

What properties would we like from such a class?

Many candidates: polynomials, trigonometric functions, continuous functions, differentiable functions, etc.

Desirable Properties

Property 1

It should be a rich class with good approximation power.

Property 2

$\mathcal{F}$ should have linear structure.

Property 3

We would like it to have inner product so that we can take projections as we have seen for linear functions, i.e. Hilbert space (complete vector space with inner product).

$$\langle f, f \rangle \geq 0$$

$$\langle f, \alpha g + \beta h \rangle = \alpha \langle f, g \rangle + \beta \langle f, h \rangle$$

Evaluation Functionals Bounded

Suppose $D_1 \mapsto f_{D_1}$ and $D_2 \mapsto f_{D_2}$

If $\|f_{D_1} - f_{D_2}\|$ is small, then f_{D_1} and f_{D_2} will make similar predictions at each point x,
i.e., $|f_{D_1}(x) - f_{D_2}(x)|$ will be small.

Evaluation Functionals Bounded

Suppose $D_1 \mapsto f_{D_1}$ and $D_2 \mapsto f_{D_2}$

If $\|f_{D_1} - f_{D_2}\|$ is small, then f_{D_1} and f_{D_2} will make similar predictions at each point x,
i.e., $|f_{D_1}(x) - f_{D_2}(x)|$ will be small.

$eval_x : \mathcal{F} \to \mathbb{R}$ given by $eval_x[f] = f(x)$

$$\sup_f \frac{|eval_x(f)|}{\|f\|} < \infty$$

Evaluation Functionals Bounded

Suppose $D_1 \mapsto f_{D_1}$ and $D_2 \mapsto f_{D_2}$

If $\|f_{D_1} - f_{D_2}\|$ is small, then f_{D_1} and f_{D_2} will make similar predictions at each point x,
i.e., $|f_{D_1}(x) - f_{D_2}(x)|$ will be small.

$$eval_x : \mathcal{F} \to \mathbb{R} \text{ given by } eval_x[f] = f(x)$$

$$\sup_f \frac{|eval_x(f)|}{\|f\|} < \infty$$

$$\sup_{x \in X} |f_{D_1}(x) - f_{D_2}(x)| \leq C \|f_{D_1} - f_{D_2}\|$$

Theorem

Any Hilbert Space where the evaluation functionals are bounded is a Reproducing Kernel Hilbert Space.

Mercer Kernel
X is a compact metric space.
$K : X \times X \to \mathbb{R}$ is a continuous **kernel**
such that
(i) $K(x, y) = K(y, x)$
(ii) for all $x_1, \ldots, x_n \in X$,

$$\mathbf{K}_{ij} = K(x_i, x_j)$$

is positive semi-definite.

Examples

$$X \subset \mathbb{R}^n : K(a,b) = e^{-\frac{||a-b||^2}{\sigma^2}}$$

$$X \subset \mathbb{R}^n : K(a,b) = (1 + a \cdot b)^d$$

$X = \{1, \ldots, k\} : K$ is $k \times k$ positive semi-definite matrix

$$X = S^1 : K(\theta, \phi) = \sum_{n=0}^{\infty} e^{-n^2 t} \sin(n\theta) \sin(n\phi)$$

Construction

1. Begin with $H_0 = \{K_x \mid x \in X\}$

2. Take finite linear combinations

   $H_1 = \{$ finite linear combinations of functions in $H_0\}$

3. Put an inner product structure

$$\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{y_j} \rangle = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j)$$

4. H_K is the completion of H_1.

Linear kernel example

$$x, y \in X = \mathbb{R}^n$$

$$K(x, y) = x \cdot y$$

$$K_x(y) = x \cdot y$$

$\sum_i \alpha_i K_{x_i}$ also a linear function

H_K is the set of linear functions

Reproducing Property

$$f(x) = \langle f, K_x \rangle$$

Therefore, by Schwarz Inequality,

$$|f(x)| \leq \|f\| \|K_x\| = \|f\| \left(K(x,x)\right)^{\frac{1}{2}} \leq K(x,x)\kappa$$

where $\kappa^2 = \sup_{x \in X} K(x,x)$.

In other words, if $\|f - g\|$ is small, then $|f(x) - g(x)|$ is small.

Another view

Let μ be a probability measure supported on X.

$$L^2(\mu) = \left\{ f \;\middle|\; \int |f|^2 d\mu < \infty \right\}$$

$L_K : L^2(\mu) \mapsto L^2(\mu)$ is an integral operator given by

$$L_K[f] = g = \int f(y) K(x, y) d\mu(y)$$

Corresponding Eigensystem

$$L_K \phi_i = \lambda_i \phi_i$$

Functions in L^2 can be written as $f = \sum_i \alpha_i \phi_i$ where $\sum_i \alpha_i^2 < \infty$.

Functions in H_K can be written as $f = \sum_i \alpha_i \phi_i$ where $\sum_i \frac{\alpha_i^2}{\lambda_i} < \infty$

Although λ_i and ϕ_i depend on the measure μ, the RKHS H_K does not.

Feature Map Interpretation

For every Mercer kernel K there exist many feature maps

$$\psi : X \to H$$

where H is a Hilbert space such that

$$K(x, y) = \langle \psi(x), \psi(y) \rangle$$

Feature Map Examples

$$\psi : X \to H_K$$

where $\psi(x) = K_x$.

Then,

$$\langle \psi(x), \psi(y) \rangle = \langle K_x, K_y \rangle = K(x, y)$$

Another Feature Map

$$\psi : X \to l_2$$

where

$$\psi(x) = (\sqrt{\lambda_1}\phi_1(x), \dots, \sqrt{\lambda_i}\phi_i(x), \dots)$$

$$\langle \psi(x), \psi(y) \rangle_{l_2} = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y) = K(x, y)$$

(Spectral theorem)