Support Distribution Machines

Barnabás Póczos, CMU
Liang Xiong, CMU
Dougal Sutherland, CMU
Jeff Schneider, CMU
Your Name, CMU

Machine Learning Lunch seminar
Carnegie Mellon University

Jan 23, 2012
bapoczos@cs.cmu.edu
Outline

Goal:
- nonparametric divergence estimation
- generalization of SVM to distributions ⇒ SDM
- applications

- Definitions and motivation
- The estimators
- Theoretical results - Consistency
- Support Distribution Machines
- Experimental results
Measuring **uncertainty** of a distribution

\[ H = - \int f(x) \log f(x) \, dx \]

\[ R_\alpha = \frac{1}{1 - \alpha} \log \int f^\alpha(x) \, dx \]

\[ T_\alpha = \frac{1}{\alpha - 1} \left( \int f^\alpha(x) \, dx - 1 \right) \]
Measuring **divergences**

Manchester United 07/08

---

**Shot Type**
- ☀ Goals
- ✧ Shots on Goal
- ○ Shots

---

**KL**

\[ KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx \]

**Tsallis**

\[ T_\alpha(p||q) = \frac{1}{\alpha - 1} \left( \int p^\alpha(x)q^{1-\alpha}(x) \, dx - 1 \right) \]

**Rényi**

\[ R_\alpha(p||q) = \frac{1}{\alpha - 1} \log \int p^\alpha(x)q^{1-\alpha}(x) \, dx \]

---

Owen Hargreaves
Rio Ferdinand
Cristiano Ronaldo

www.juhokim.com/projects.php
Measuring **dependence**

Mutual Information = dependence between random variables

Divergence between \( p = p(x_1, \ldots, x_d) \) and \( q = \prod_{i=1}^{d} p_i(x_i) \)

\[
KL(p \| \prod p_i) \quad T_\alpha(p \| \prod p_i) \quad R_\alpha(p \| \prod p_i)
\]
Who cares about dependence?

Applications: *Too many to list*…

independence tests, system identification, optimal experiment design, analysis of stock markets, feature selection, boosting, clustering, image registration,

information theory, information geometry, prediction of protein structure, drug design, fMRI data processing, microarray data processing, independent component analysis,

A. Fernandes & G. Gloor: **Mutual information is critically dependent on prior assumptions:** *would the correct estimate of mutual information please identify itself?*

*BIOINFORMATICS* Vol. 26 no. 9 2010, pages 1135–1139
How should we estimate them?

- Naïve plug-in approach using density estimation
  - density estimators
    - histogram
    - kernel density estimation
    - k-nearest neighbors [D. Loftsgaarden & C. Quesenberry. 1965.]

\[
KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx
\]

\[
T_\alpha(p\|q) = \frac{1}{\alpha - 1} \left( \int p^\alpha(x) q^{1-\alpha}(x) \, dx - 1 \right)
\]

\[
R_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \int p^\alpha(x) q^{1-\alpha}(x) \, dx
\]

Density: nuisance parameter
Density estimation: difficult

How can we estimate them directly?
The Estimator

$X_{1:n} = \{X_1, \ldots, X_n\} \sim p$
$Y_{1:n} = \{Y_1, \ldots, Y_m\} \sim q$

$k = 2.$

$D_\alpha(p\|q) \doteq \int_M p^\alpha(x) q^{1-\alpha}(x) \, dx.$

$k \geq 1$, fixed.

$\rho_k(i) :$ the distance of the $k$-th nearest neighbor of $X_i$ in $X_{1:n}$
$\nu_k(i) :$ the distance of the $k$-th nearest neighbor of $X_i$ in $Y_{1:m}$

$$\hat{D}_\alpha(X_{1:n}\|Y_{1:m}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n-1)\rho_k^d(i)}{m\nu_k^d(i)} \right)^{1-\alpha} \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$$

First direct estimator for $\alpha$ divergences
**Special cases: MI estimation**

MI: measure the dependence between the random variables.

\[
I(X^1, \ldots, X^d) = KL(p(x^1, \ldots, x^d) \| p_1(x^1) \cdots p_d(x^d)) \\
I_\alpha(X^1, \ldots, X^d) = D_\alpha(p(x^1, \ldots, x^d) \| p_1(x^1) \cdots p_d(x^d))
\]

\[X_{1:n} \overset{\sim}{=} (X_1, \ldots, X_n) \text{ i.i.d. sample } \sim p, \ X_j \in \mathbb{R}^d\]

\[X^1 = \{X^1_1, \ldots, X^1_n\} \quad \sim p \quad \Rightarrow \quad Y^1 = \{X^{\sigma_1(1)}_1, \ldots, X^{\sigma_1(n)}_1\} \quad \sim \prod_{i=1}^d p_i = q\]

\[X^i = \{X^i_1, \ldots, X^i_n\} \quad \Rightarrow \quad Y^i = \{X^{\sigma_i(1)}_i, \ldots, X^{\sigma_i(n)}_i\}\]

\[X^d = \{X^d_1, \ldots, X^d_n\} \quad Y^d = \{X^{\sigma_d(1)}_d, \ldots, X^{\sigma_d(n)}_d\}\]
Theoretical Results
(Asymptotically unbiased, L2 consistent)
Notations: Stochastic convergence

Let \( \{Z, Z_1, Z_2, \ldots \} \) be a sequence of random variables converge in distribution, or converge weakly, or converge in law

\[
Z_n \xrightarrow{d} Z, \quad F_n \xrightarrow{w} F
\]

\[
\lim_{n \to \infty} F_n(z) = F(z), \quad \forall z \in \mathbb{R} \text{ at which } F \text{ is continuous}
\]

converge in probability \( Z_n \overset{p}{\to} Z \)

\[
\forall \varepsilon > 0 \quad \lim_{n \to \infty} \Pr \left( |Z_n - Z| \geq \varepsilon \right) = 0.
\]

Almost surely convergence \( Z_n \xrightarrow{a.s.} Z \)

\[
\Pr \left( \omega \in \Omega : \lim_{n \to \infty} Z_n(\omega) = Z(\omega) \right) = 1.
\]

convergence in p-th mean, convergence in L_p norm \( Z_n \overset{L_p}{\to} Z \)

\[
\lim_{n \to \infty} \mathbb{E} \left[ |Z_n - Z|^p \right] = 0
\]
k-NN density estimators

What is a density? [Lebesgue, 1910]

Density: Let \( P \) be a probability measure.

\[
f(Z = z) = \lim_{r \to 0} \frac{P(Z \in B(z, r))}{Vol(B(z, r))}, \text{ for a.a. } z
\]

Estimate \( P(Z \in B(z, r)) \) with the empirical distribution!

_The estimation is tricky, because r should converge to 0, but if it converges too fast, that is not good either…_
k-NN density estimators

The estimation is tricky, because \( r \) should converge to 0, but if it converges too fast, that is not good either…

Choose \( r \) adaptively: \( r = \rho_k(i) \! \\
\rho_k(i) : \text{the distance of the } k\text{-th nearest neighbor of } X_i \text{ in } X_{1:n} \)

The estimated density at \( X_i = x \):

\[
\hat{f}(x) = \frac{k/(n-1)}{\text{Vol}(B(x, \rho_k))}
\]

How good is this estimation?
k-NN density estimators

Definition: \[ Z_n \rightarrow_p Z \Leftrightarrow \forall \varepsilon > 0 : \lim_{n \to \infty} \Pr \left( |Z_n - Z| \geq \varepsilon \right) = 0. \]

Theorem:

If \[ \lim_{n \to \infty} k(n) = \infty \]
and \[ \lim_{n \to \infty} n/k(n) = \infty \]
\[ \Rightarrow \text{ then for almost all } x \hat{p}_k(n)(x) \to^p p(x). \]

If \[ \lim_{n \to \infty} k(n)/\log(n) = \infty \]
and \[ \lim_{n \to \infty} n/k(n) = \infty \]
\[ \Rightarrow \text{ then } \lim_{n \to \infty} \sup_{x} |\hat{p}_k(n)(x) - p(x)| = 0. \text{ (a.s.)} \]

Usually \( k(n) \sim \sqrt{n} \)

\[ \hat{p}_k(X_i) = \frac{k/(n-1)}{\text{Vol}(B(X_i, \rho_k))} \]

\[ \hat{q}_k(X_i) = \frac{k/m}{\text{Vol}(B(X_i, \nu_k))} \]

We will keep \( k \) fixed!
The Estimator, revisited

\[ D_\alpha(p\|q) = \int_M \left( \frac{q(x)}{p(x)} \right)^{1-\alpha} p(x) \, dx. \]

\[ \hat{p}_k(i) = \frac{k/(n-1)}{c\rho^d_k(i)}, \quad \hat{q}_k(i) = \frac{k/m}{c\nu^d_k(i)} \]

\[ \Rightarrow \tilde{D}_\alpha(p\|q) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n-1)\rho^d_k(i)}{m\nu^d_k(i)} \right)^{1-\alpha} \]

It has a multiplicative bias, but this bias (asymptotically) is independent of \( p \) and \( q \)!

\[ \tilde{D}_\alpha(X_{1:n}\|Y_{1:m}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n-1)\rho^d_k(i)}{m\nu^d_k(i)} \right)^{1-\alpha} \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)} \]
Main Theorems

Asymptotically unbiased

\[ \lim_{n,m \to \infty} \mathbb{E} \left[ \hat{D}_\alpha(X_{1:n} \| Y_{1:m}) \right] = D_\alpha(p \| q) \]

$L_2$ consistent

\[ \lim_{n,m \to \infty} \mathbb{E} \left[ \left( \hat{D}_\alpha(X_{1:n} \| Y_{1:m}) - D_\alpha(p \| q) \right)^2 \right] = 0 \]

Sufficient conditions

- $p$, $q$ bounded away from zero
- $p$, $q$ bounded above
- $p$, $q$ uniformly continuous densities
- $-k < \min(1-\alpha,\alpha-1) < \max(1-\alpha,\alpha-1) < k$
- conditions on the domain

allowed

not allowed
The distribution of “normalized” k-NN distances

**Normalized k-NN distances** Let \( \zeta_{n,1} \doteq (n - 1)\rho_k^d(1) \)

Let \( F_{n,x}(u) \doteq \Pr(\zeta_{n,1} < u | X_1 = x) \) be its conditional cdf.

**Theorem** ([Normalized k-NN distances have “binomial” cdfs])

\[
F_{n,x}(u) = 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-j},
\]

where \( R_n(u) \doteq \left( \frac{u}{n-1} \right)^{1/d}, \) and \( P_{n,u,x} \doteq \int_{\mathcal{M} \cap B(x,R_n(u))} p(t) dt \)
“normalized” k-NN distances

Proof \[ \zeta_{n,1} \doteq (n - 1) \rho^d_k(1) \]

\[ F_{n,x}(u) \doteq \Pr(\zeta_{n,1} < u | X_1 = x) \]

\[ = \Pr((n - 1) \rho^d_k(1) < u | X_1 = x) \]

\[ = \Pr \left( \rho_k(1) < \left( \frac{u}{n - 1} \right)^{1/d} | X_1 = x \right) \]

\[ = \Pr(\rho_k(1) < R_n(u) | X_1 = x) \]

\[ = \Pr(k \text{ elements or more from } \{X_2, \ldots, X_n\} \in B(x, R_n(u)) \cap \mathcal{M} | X_1 = x) \]

\[ = \Pr(k \text{ elements or more from } \{X_2, \ldots, X_n\} \in B(x, R_n(u)) \cap \mathcal{M}) \]

\[ = \sum_{j=k}^{n-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j} \text{ binomial distribution} \]

\[ = 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j}. \]

where \[ R_n(u) \doteq \left( \frac{u}{n-1} \right)^{1/d}, \text{ and } P_{n,u,x} \doteq \int_{\mathcal{M} \cap B(x,R_n(u))} p(t) dt \]
Erlang distribution

Theorem: \( F_{n,x} \rightarrow^w F_x \) (for almost all \( x \in M \)),

where \( F_x(u) = 1 - \exp(-\lambda u) \sum_{j=0}^{k-1} \frac{(\lambda u)^j}{j!} \)

is the cdf of the Erlang distribution with \( \lambda = \bar{c}p(x) \), \( \bar{c} \) is the volume of a unit ball.

Normalized k-NN distances converge to Erlang distrib.

\( k \): shape, \( 1/\lambda \): scale

mean = \( k/\lambda = k/(\bar{c}p(x)) \)

var = \( k/\lambda^2 = k/(\bar{c}^2p^2(x)) \)

\( \zeta_{n,1}\bar{c}/k \equiv \bar{c}(n-1)\rho_k^d(1)/k \rightarrow_d \begin{cases} 1/p(x) \text{ mean} \\ 1/kp^2(x) \text{ var} \end{cases} \)
Erlang distribution

Theorem:  \( F_{n,x}(u) \to F_x(u) \ \forall u, \) (for a.a. \( x \))

\[ F_x(u) = 1 - \sum_{j=0}^{k-1} \frac{(\lambda u)^j}{j!} e^{-\lambda u} \]

\( \lambda = \bar{c} p(x) \), \( \bar{c} \) is the volume of a unit ball.

Proof:  \( p(X = x) = \lim_{r \to 0} \frac{\mathbb{P}(X \in B(x,r))}{\text{Vol}(B(x,r))} \)

\[ \Rightarrow \text{If } n > n_0, \text{ then } p(x) - \delta < \frac{\int_{B(x,R_n(u))} p(t) \ dt}{\text{Vol}(B(x,R_n(u)))} = \frac{P_{n,u,x}}{\bar{c}u/(n-1)} < p(x) + \delta \]

\[ \Rightarrow \text{If } n > n_0, \text{ then } \frac{(p(x) - \delta) \bar{c}u}{(n-1)} < P_{n,u,x} < \frac{(p(x) + \delta) \bar{c}u}{(n-1)} \]

\[ F_{n,x}(u) = 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j} \]

\[ \geq 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} \left( \frac{(p(x) + \delta) \bar{c}u}{n-1} \right)^j \left( 1 - \frac{(p(x) - \delta) \bar{c}u}{n-1} \right)^{n-1-j} \]

\[ \to F_x(u) \]

The other direction goes similarly.
**Key Lemma:** Moments of normalized k-NN distances in the limit

Let $\zeta_{n,1} \triangleq (n-1)\rho_k^d(1)$ be normalized k-NN distances.

Let $F_{n,x}(u) \triangleq \Pr(\zeta_{n,1} < u | X_1 = x)$ be its conditional cdf.

Let $\xi_{n,x} \sim F_{n,x}$, $\xi_x \sim F_x$ be random variables.

We already know that $\xi_{n,x} \to_d \xi_x \Rightarrow \xi_{n,x}^\gamma \to_d \xi_x^\gamma$, $\forall \gamma$.

If from this it follows that $\mathbb{E}[\xi_{n,x}^\gamma] \to \mathbb{E}[\xi_x^\gamma]$, then

$$\lim_{n \to \infty} \mathbb{E}\left[(n-1)^\gamma \rho_k^d(1) | X_1 = x\right] = \lim_{n \to \infty} \mathbb{E}\left[\xi_{n,x}^\gamma\right]$$

$$= \mathbb{E}[\xi_x^\gamma] = \int_0^\infty u^\gamma f_x(u) \, dx = (\bar{c}p(x))^{-\gamma} \frac{\Gamma(k + \gamma)}{\Gamma(k)},$$

Erlang $\gamma$-th moment

Erlang density

Similarly,

$$\lim_{m \to \infty} \mathbb{E}\left[m^\gamma \nu_k^d(1) | X_1 = x\right] = (\bar{c}q(x))^{-\gamma} \frac{\Gamma(k + \gamma)}{\Gamma(k)}.$$
### Asymptotically unbiased

The estimator

\[
\widehat{D}_\alpha(X_{1:n} \mid Y_{1:m}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n - 1) \rho^d_k(i)}{m \nu^d_k(i)} \right)^{1-\alpha} \frac{\Gamma(k)^2}{\Gamma(k - \alpha + 1) \Gamma(k + \alpha - 1)}
\]

We want to prove that the estimator is asymptotically unbiased:

\[
D_\alpha(p \mid q) \frac{\Gamma(k - \alpha + 1) \Gamma(k + \alpha - 1)}{\Gamma(k)^2} = \lim_{n,m \to \infty} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n - 1) \rho^d_k(i)}{m \nu^d_k(i)} \right)^{1-\alpha} \right].
\]

The r.h.s. can be rewritten as

\[
\mathbb{E} \left[ \frac{(n - 1)(1-\alpha) \rho^d_{k^d}(1-\alpha)}{m(1-\alpha) \nu^d_{k}(1-\alpha)}(1) \right] \neq \frac{\mathbb{E} \left[ (n - 1)(1-\alpha) \rho^d_{k^d}(1-\alpha)(1) \right]}{\mathbb{E} \left[ m(1-\alpha) \nu^d_{k}(1-\alpha)(1) \right]} \leftarrow \mathbb{E} \left[ \frac{X}{Y} \right] \neq \frac{\mathbb{E} [X]}{\mathbb{E} [Y]}
\]

\[
\neq \mathbb{E} \left[ (n - 1)(1-\alpha) \rho^d_{k^d}(1-\alpha)(1) \right] \mathbb{E} \left[ \frac{1}{m(1-\alpha) \nu^d_{k}(1-\alpha)(1)} \right] \leftarrow \text{dependent}
\]

\[
\mathbb{E} \left[ \frac{X}{Y} \right] \neq \mathbb{E} [X] \mathbb{E} \left[ \frac{1}{Y} \right]
\]
Asymptotically unbiased

The r.h.s. can be rewritten as

\[
\lim_{n,m \to \infty} \mathbb{E}_{X_1 \sim p} \left[ \mathbb{E} \left[ (n - 1)^{1-\alpha} \rho_k^{d(1-\alpha)}(1) \bigg| X_1 \right] \mathbb{E} \left[ m^{\alpha-1} \nu_k^{d(\alpha-1)}(1) \bigg| X_1 \right] \right].
\]

If we could move the limit inside the expectations...

\[
= \mathbb{E}_{X_1 \sim p} \left[ \lim_{n \to \infty} \mathbb{E} \left[ (n - 1)^{1-\alpha} \rho_k^{d(1-\alpha)}(1) \bigg| X_1 \right] \lim_{m \to \infty} \mathbb{E} \left[ m^{\alpha-1} \nu_k^{d(\alpha-1)}(1) \bigg| X_1 \right] \right]
\]

\[
= \mathbb{E}_{X_1 \sim p} \left[ \lim_{n \to \infty} (n - 1)^{1-\alpha} \rho_k^{d(1-\alpha)}(1) \bigg| X_1 \right] \mathbb{E} \left[ \lim_{m \to \infty} m^{\alpha-1} \nu_k^{d(\alpha-1)}(1) \bigg| X_1 \right] \right]
\]

\[
= \mathbb{E}_{X_1 \sim p} \left[ \frac{(\bar{c}p(X_1))^{(\alpha-1)}}{(\bar{c}q(X_1))^{(\alpha-1)}} \frac{\Gamma(k - \alpha + 1) \Gamma(k + \alpha - 1)}{\Gamma(k)} \frac{\Gamma(k + \alpha - 1)}{\Gamma(k)} \right]
\]

\[
= D_\alpha(p||q) \frac{\Gamma(k - \alpha + 1) \Gamma(k + \alpha - 1)}{\Gamma(k)} \frac{\Gamma(k + \alpha - 1)}{\Gamma(k)}
\]

All what we need is \( \{\xi_n \to_d \xi\} \Rightarrow \{\mathbb{E}[\xi_n^\gamma] \to \mathbb{E}[\xi^\gamma]\} \)
A little problem...

\[ \{ \xi_n \to_d \xi \} \Rightarrow \{ \mathbb{E}[\xi_n] \to \mathbb{E}[\xi] \} \]

\[ \mathbb{E}[\xi] < \infty, \text{ but } \mathbb{E}[\xi_1] = \infty, \mathbb{E}[\xi_2] = \infty, \ldots, \mathbb{E}[\xi_n] = \infty, \ldots \]

Solutions:

Asymptotically uniformly integrability…

\[ \lim_{\beta \to \infty} \limsup_{n \to \infty} \int_{|u| \geq \beta} |u| f_n(u) \, du = 0 \quad \text{then} \quad \lim_{n \to \infty} \mathbb{E}[\xi_n] = \mathbb{E}[\xi]. \]

Upper bounding

Let \( \xi_n \to_d \xi \), \( 0 \leq \xi_n, \xi \), and \( \gamma \in \mathbb{R} \). If there exists an \( \varepsilon > 0 \) such that \( \limsup_{n \to \infty} \mathbb{E}[\xi_n^{\gamma(1+\varepsilon)}] < \infty \), then \( \lim_{n \to \infty} \mathbb{E}[\xi_n^\gamma] = \mathbb{E}[\xi^\gamma] \).
Be careful, some mistakes are easy to make...

We want:
\[ \{ F_n(u) \to F(u) \ \forall u \} \Rightarrow \left\{ \int_0^\infty u \, dF_n(u) \to \int_0^\infty u \, dF(u) \right\} \]

Helly–Bray theorem
\[ \int_\mathbb{R} g(u) \, dF_n(u) \to \int_\mathbb{R} g(u) \, dF(u) \]
for each bounded, continuous function \( g : \mathbb{R} \to \mathbb{R} \),

[Annals of Statistics]
Some mistakes ...

We want:
\[ \{F_n(u) \to F(u) \ \forall u\} \Rightarrow \left\{ \int_0^\infty u \, dF_n(u) \to \int_0^\infty u \, dF(u) \right\} \]

Enough:
There exists an \( \varepsilon > 0 \) such that \( \limsup_{n \to \infty} \mathbb{E} \left[ \xi_n^{\gamma(1+\varepsilon)} \right] < \infty \).

\[ \text{Fatou lemma:} \]
\[ \limsup_{n \to \infty} \mathbb{E} \left[ \xi_n^{\gamma(1+\varepsilon)} \right] \leq \mathbb{E} \left[ \limsup_{n \to \infty} \xi_n^{\gamma(1+\varepsilon)} \right] < \infty \]

\[ \text{Fatou lemma:} \]
\[ \gamma(1+\varepsilon) \text{ moment of an Erlang variable} < \infty \]

Extensions

**$L_2$ divergence:**

\[
L(p||q) \doteq \left( \int_M (p(x) - q(x))^2 \, dx \right)^{1/2}
\]

\[
\hat{L}^2(X_1:N||Y_1:M) \doteq \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{k-1}{(N-1) \rho_{k}^d(X_n)} - \frac{2(k-1)}{M \rho_{k}^d(X_n)} + \frac{(N-1) \rho_{k}^d(X_n)(k-2)(k-1)}{(M \rho_{k}^d(X_n))^2} \right]
\]

where $k - 2 > 0$.

**Conditional Renyi Mutual Information:**

\[
I_\alpha(X, Y|Z) \doteq \int p_Z(z) D_\alpha(p(X, Y|Z = z)||p(X|Z = z)p(Y|Z = z)|Z = z)
\]

\[
\hat{I}_\alpha = \frac{1}{\alpha - 1} \log \frac{1}{N} \sum_{n=1}^{N} \frac{(c_{xyz})^{(1-\alpha)} \rho_{xyz}^{(1-\alpha)}(X_n; Y_n; Z_n)}{(c_{xz})^{(1-\alpha)} \rho_{xz}^{(1-\alpha)}(X_n; Z_n)} \frac{(c_{yz})^{(1-\alpha)} \rho_{yz}^{(1-\alpha)}(Y_n; Z_n)}{(c_{yz})^{(1-\alpha)} \rho_{yz}^{(1-\alpha)}(Y_n; Z_n)} B^2
\]

where $B^2 = \frac{\Gamma^4(k)}{\Gamma^2(k-\alpha+1) \Gamma^2(k+\alpha-1)}$. 

\[
\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, dt
\]
Numerical Experiments
Beta distributions, 
$\alpha=0.4$, 5 independent experiments
Results for divergence estimation

2D Normal
Results for MI estimation

rotated uniform distribution $\alpha = 0.999$

(a) Observations from $f$ and $g$
Result for MI estimation

2D Gaussian
Support Distribution Machines
Support Distribution Machines

Goal: generalize kernel machines to the space of distributions.

Classify the following distributions (sample sets) using SVMs
Distribution Classification

We have $T$ sample sets, $(X_1, \ldots, X_T)$. [Training data] $X_t$ consists of $m_t$ sample points i.i.d sampled from $p_t$.

$\{X_{t,1}, \ldots, X_{t,m_t}\} = X_t \sim p_t$. $X_t$ has class label $Y_t \in \{-1, +1\}$.

What is the class label $Y$ of $\{X_1, \ldots, X_m\} = X \sim p$?

Use RKHS based SVMs to solve this problem!

$K_{ij} = \langle \phi(p_i), \phi(p_j) \rangle_K = K(p_i, p_j)$

Dual form of SVM:

$\hat{\alpha} = \text{arg max}_{\alpha \in \mathbb{R}^T} \sum_{i=1}^{T} \alpha_i - \frac{1}{2} \sum_{i,j}^{T} \alpha_i \alpha_j y_i y_j K_{ij}$, subject to $\sum_i \alpha_i y_i = 0$, $0 \leq \alpha_i \leq C$.

$Y = \text{sign}(\sum_{i=1}^{T} \hat{\alpha}_i y_i K(p_i, p)) \in \{-1, +1\}$

Problems: We do not know $p_i$, $p$, $K(p_i, p_j)$, or $K(p_i, p)$...
Kernel Estimation

Linear kernel: \( K(p, q) = \int pq \)

Polynomial kernel: \( K(p, q) = (\int pq + c)^s \)

Gaussian kernel: \( K(p, q) = \exp\left(-\frac{1}{2\sigma^2}\mu^2(p, q)\right), \mu(p, q) = (\int p^2 + q^2 - 2pq)^{1/2} \).

We only need to estimate \( \int p^\alpha q^\beta \) terms.

We can also try to use other \( \mu(p, q) \) divergences, e.g. Rényi ...

Problems???

The \( \{\widehat{K}_{i,j}\}_{ij} \) Gram matrix might not be PSD!

Solution: make it symmetric, and project it to the cone of PSD matrices!
Numerical Experiments
Noisy USPS Dataset Classification

Sample size of each object: 300
100 training and 100 test data

Results:
SVM on raw images with degree 3 polynomial kernel: 67% accuracy
SDM on the 2D distributions with the same kernel: 91% accuracy
Multidimensional Scaling of USPS Data

Nonlinear embedding with MDS into 2D.
10 instances from letters 1,2,3,4.

Raw images using Euclidean distance

Estimated Euclidean distance between the distributions
Image to Distribution Transformation
Image to Distribution Transformation

Each image is represented by a set of 128 dimensional vectors. Compressed by PCA to 18 – 53 dim.
Object Classification
ETH-80 [Leibe and Schiele, 2003]

• 8 categories
• BoW: 88.9%
• NPR-0.9: 90.1%

Each image is represented by 576 18D points. Significant difference. 16 runs.
Scenes Classification
[Oliva and Torralba, 2001]

- 8 categories
- Best published: **91.57%** (Qin and Yung, ICMV 2010)
- NPR-0.99: **92.3%**

Image represented by 1815 53D points. 16 random runs.
Sport Events Classification
[Li and Fei Fei, 2007]

- 8 categories
- Best published: **86.7%**
  (Zhang et al, CVPR 2011)
- NPR-0.9: **87.1%**

Image represented by 295-1542 53D points. 16 random runs.
Turbulence data classification

- JHU simulated fluid flow through time
- **Goal**: find vortices.
- 11 positive, 20 negative examples

- **Results**: Leave one out cross validation: 97% accuracy

**Velocity distributions:**

Positive (vortex)  |  negative  |  negative

Vorticity scores
Finding vortices

Classification probabilities
Find interesting phenomena in turbulence data

Anomaly detection with 1-class SDM

Anomaly scores
Distribution Regression
with Scalar Response

Learn $\mathcal{F}(\mathbf{p}) \in \mathbb{R}$ functional of the density!

Skewness of Beta

Entropy of Gaussian
Local Linear Embedding of Distributions

Nonlinear dimensionality reduction into 2D
Local Linear Embedding of Distributions

72 rotated COIL froggies

Edge detected COIL froggy

Euclidean distance between images

Euclidean distance between distributions
Conclusions

- Many ML problems can be solved efficiently on the space of distributions.
- Kernel methods can be generalized to sample sets and distributions.
- No need to estimate densities!

If you need to estimate divergences, then use me!

\[
\hat{D}_\alpha(X_1:n \| Y_1:m) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(n - 1) \rho_k^d(i)}{mv_k^d(i)} \right)^{1-\alpha} \frac{\Gamma(k)^2}{\Gamma(k - \alpha + 1) \Gamma(k + \alpha - 1)}
\]

Thanks for your attention! 😊