Kernel Methods

Lecture 4: Maximum Mean Discrepancy

Thanks to Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, Jiayuan Huang, Arthur Gretton

Alexander J. Smola

Statistical Machine Learning Program
Canberra, ACT 0200 Australia
Alex.Smola@nicta.com.au

Machine Learning Summer School, Taiwan 2006

Course Overview

1. Estimation in exponential families
   - Maximum Likelihood and Priors
   - Clifford Hammersley decomposition
2. Applications
   - Conditional distributions and kernels
   - Classification, Regression, Conditional random fields
3. Inference and convex duality
   - Maximum entropy inference
   - Approximate moment matching
4. Maximum mean discrepancy
   - Means in feature space, Covariate shift correction
5. Hilbert-Schmidt independence criterion
   - Covariance in feature space
   - ICA, Feature selection

Course Overview

1. Estimation in exponential families
   - Maximum Likelihood and Priors
   - Clifford Hammersley decomposition
2. Applications
   - Conditional distributions and kernels
   - Classification, Regression, Conditional random fields
3. Inference and convex duality
   - Maximum entropy inference
   - Approximate moment matching
4. Maximum mean discrepancy
   - Means in feature space, Covariate shift correction
5. Hilbert-Schmidt independence criterion
   - Covariance in feature space
   - ICA, Feature selection

Course Overview

1. Estimation in exponential families
   - Maximum Likelihood and Priors
   - Clifford Hammersley decomposition
2. Applications
   - Conditional distributions and kernels
   - Classification, Regression, Conditional random fields
3. Inference and convex duality
   - Maximum entropy inference
   - Approximate moment matching
4. Maximum mean discrepancy
   - Means in feature space, Covariate shift correction
5. Hilbert-Schmidt independence criterion
   - Covariance in feature space
   - ICA, Feature selection

Outline

1. **Two Sample Problem**
   - Direct Solution
   - Kolmogorov Smirnov Test
   - Reproducing Kernel Hilbert Spaces
   - Test Statistics
2. Data Integration
   - Problem Definition
   - Examples
3. Attribute Matching
   - Basic Problem
   - Linear Assignment Problem
4. Sample Bias Correction
   - Sample Reweighting
   - Quadratic Program and Consistency
   - Experiments

Two Sample Problem

Setting

Given $X := \{x_1, \ldots, x_m\} \sim p$ and $Y := \{y_1, \ldots, y_n\} \sim q$, test whether $p = q$.

Applications

- Cross platform compatibility of microarrays
  Need to know whether distributions are the same.
- Database schema matching
  Need to know which coordinates match.
- Sample bias correction
  Need to know how to reweight data.
- Feature selection
  Need features which make distributions most different.
- Parameter estimation
  Reduce two-sample to one-sample test.

Two Sample Problem

Setting

Given $X := \{x_1, \ldots, x_m\} \sim p$ and $Y := \{y_1, \ldots, y_n\} \sim q$, test whether $p = q$.

Applications

- Cross platform compatibility of microarrays
  Need to know whether distributions are the same.
- Database schema matching
  Need to know which coordinates match.
- Sample bias correction
  Need to know how to reweight data.
- Feature selection
  Need features which make distributions most different.
- Parameter estimation
  Reduce two-sample to one-sample test.

Indirect Solution

Plug in estimators

Estimate $\hat{p}$ and $\hat{q}$ and compute $D(\hat{p}, \hat{q})$.

Parzen Windows and L_2 distance

$$\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m}\kappa(x_i, x) \text{ and } \hat{q}(y) = \frac{1}{n}\sum_{i=1}^{n}\kappa(y_i, y)$$

Computing squared L_2 distance between $\hat{p}$ and $\hat{q}$ yields

$$\|\hat{p} - \hat{q}\|_2^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)$$

where $k(x, x') = \int \kappa(x, t)\kappa(x', t)dt$.

Indirect Solution

Plug in estimators

Estimate $\hat{p}$ and $\hat{q}$ and compute $D(\hat{p}, \hat{q})$.

Parzen Windows and L_2 distance

$$\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m}\kappa(x_i, x) \text{ and } \hat{q}(y) = \frac{1}{n}\sum_{i=1}^{n}\kappa(y_i, y)$$

Computing squared L_2 distance between $\hat{p}$ and $\hat{q}$ yields

$$\|\hat{p} - \hat{q}\|_2^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)$$

where $k(x, x') = \int \kappa(x, t)\kappa(x', t)dt$.

Problems

- Curse of dimensionality when estimating $\hat{p}$ and $\hat{q}$
- Statistical analysis (multi-stage procedure).
- What to do on strings, images, structured data?
- This quantity is biased (even for $p = q$ its expected value does not vanish).

Direct Solution

Key Idea

Avoid density estimator, use **means** directly.

Maximum Mean Discrepancy (Fortet and Mourier, 1953)

$$D(p, q, \mathcal{F}) := \sup_{f \in \mathcal{F}} \mathbf{E}_p [f(x)] - \mathbf{E}_q [f(y)]$$

Theorem (via Dudley, 1984)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = C^0(\mathcal{X})$ is the space of continuous, bounded, functions on $\mathcal{X}$.

Theorem (via Steinwart, 2001; Smola et al., 2006)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = \{f \mid \|f\|_{\mathcal{H}} \leq 1\}$ is a unit ball in a Reproducing Kernel Hilbert Space, provided that $\mathcal{H}$ is universal.

Direct Solution

Key Idea

Avoid density estimator, use **means** directly.

Maximum Mean Discrepancy (Fortet and Mourier, 1953)

$$D(p, q, \mathcal{F}) := \sup_{f \in \mathcal{F}} \mathbf{E}_p\left[f(x)\right] - \mathbf{E}_q\left[f(y)\right]$$

Theorem (via Dudley, 1984)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = C^0(\mathcal{X})$ is the space of continuous, bounded, functions on $\mathcal{X}$.

Theorem (via Steinwart, 2001; Smola et al., 2006)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = \{f \mid \|f\|_{\mathcal{H}} \leq 1\}$ is a unit ball in a Reproducing Kernel Hilbert Space, provided that $\mathcal{H}$ is universal.

Direct Solution

Key Idea

Avoid density estimator, use **means** directly.

Maximum Mean Discrepancy (Fortet and Mourier, 1953)

$$D(p, q, \mathcal{F}) := \sup_{f \in \mathcal{F}} \mathbf{E}_p\left[f(x)\right] - \mathbf{E}_q\left[f(y)\right]$$

Theorem (via Dudley, 1984)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = C^0(\mathcal{X})$ is the space of continuous, bounded, functions on $\mathcal{X}$.

Theorem (via Steinwart, 2001; Smola et al., 2006)

$D(p, q, \mathcal{F}) = 0$ iff $p = q$, when $\mathcal{F} = \{f | \, \|f\|_{\mathcal{H}} \leq 1\}$ is a unit ball in a Reproducing Kernel Hilbert Space, provided that $\mathcal{H}$ is universal.

Direct Solution

Proof.

- If $p = q$ it is clear that $D(p, q, \mathcal{F}) = 0$ for any $\mathcal{F}$.
- If $p \neq q$ there exists some $f \in C^0(\mathcal{X})$ such that

$$\mathbf{E}_p[f] - \mathbf{E}_q[f] = \epsilon > 0$$

- Since $\mathcal{H}$ is universal, we can find some f^* such that $\|f - f^*\|_\infty \leq \frac{\epsilon}{2}$.
- Rescale f^* to fit into unit ball.

$\square$

Goals

- Empirical estimate for $D(p, q, \mathcal{F})$.
- Convergence guarantees.

Direct Solution

Proof.

- If $p = q$ it is clear that $D(p, q, \mathcal{F}) = 0$ for any $\mathcal{F}$.
- If $p \neq q$ there exists some $f \in C^0(\mathcal{X})$ such that

$$\mathbf{E}_p[f] - \mathbf{E}_q[f] = \epsilon > 0$$

- Since $\mathcal{H}$ is universal, we can find some f^* such that $\|f - f^*\|_\infty \leq \frac{\epsilon}{2}$.
- Rescale f^* to fit into unit ball.

$\square$

Goals

- Empirical estimate for $D(p, q, \mathcal{F})$.
- Convergence guarantees.

Kolmogorov Smirnov Statistic

Function Class

- Real-valued in one dimension, $\mathcal{X} = \mathbb{R}$
- $\mathcal{F}$ are all functions with **total variation** less than 1.
- **Key:** $\mathcal{F}$ is absolute convex hull of $\xi_{(-\infty,t]}(x)$ for $t \in \mathbb{R}$.

Optimization Problem

$$\sup_{f \in \mathcal{F}} \mathbf{E}_p \left[f(x) \right] - \mathbf{E}_q \left[f(y) \right] =$$

$$\sup_{t \in \mathbb{R}} \left| \mathbf{E}_p \left[\xi_{(-\infty,t]}(x) \right] - \mathbf{E}_q \left[\xi_{(-\infty,t]}(y) \right] \right| = \| F_p - F_q \|_\infty$$

Estimation

- Use empirical estimates of F_p and F_q.
- Use Glivenko-Cantelli to obtain statistic.

Hilbert Space Setting

Function Class

- Reproducing Kernel Hilbert Space $\mathcal{H}$ with kernel k.
- Evaluation functionals

$$f(x) = \langle k(x, \cdot), f \rangle .$$

- Computing means via linearity

$$\mathbf{E}_p [f(x)] = \mathbf{E}_p [\langle k(x, \cdot), f \rangle] = \left\langle \underbrace{\mathbf{E}_p [k(x, \cdot)]}_{:=\mu_p}, f \right\rangle$$

$$\frac{1}{m} \sum_{i=1}^{m} f(x_i) = \frac{1}{m} \sum_{i=1}^{m} \langle k(x_i, \cdot), f \rangle = \underbrace{\left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle}_{:=\mu_X}$$

- Computing means via $\langle \mu_p, f \rangle$ and $\langle \mu_X, f \rangle$.

Hilbert Space Setting

Optimization Problem

$$\sup_{\|f\|\leq 1} \mathbf{E}_p\left[f(x)\right] - \mathbf{E}_q\left[f(y)\right] = \sup_{\|f\|\leq 1} \langle \mu_p - \mu_q, f \rangle = \|\mu_p - \mu_q\|_{\mathcal{H}}$$

Kernels

$$\begin{aligned}\|\mu_p - \mu_q\|_{\mathcal{H}}^2 &= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle \\ &= \mathbf{E}_{p,p} \langle k(x,\cdot), k(x',\cdot) \rangle - 2\mathbf{E}_{p,q} \langle k(x,\cdot), k(y,\cdot) \rangle \\ &\quad + \mathbf{E}_{q,q} \langle k(y,\cdot), k(y',\cdot) \rangle \\ &= \mathbf{E}_{p,p} k(x,x') - 2\mathbf{E}_{p,q} k(x,y) + \mathbf{E}_{q,q} k(y,y')\end{aligned}$$

Hilbert Space Setting

Optimization Problem

$$\sup_{\|f\|\leq 1} \mathbf{E}_p [f(x)] - \mathbf{E}_q [f(y)] = \sup_{\|f\|\leq 1} \langle \mu_p - \mu_q, f \rangle = \|\mu_p - \mu_q\|_{\mathcal{H}}$$

Kernels

$$\begin{aligned}\|\mu_p - \mu_q\|_{\mathcal{H}}^2 &= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle \\ &= \mathbf{E}_{p,p} \langle k(x,\cdot), k(x',\cdot) \rangle - 2\mathbf{E}_{p,q} \langle k(x,\cdot), k(y,\cdot) \rangle \\ &\quad + \mathbf{E}_{q,q} \langle k(y,\cdot), k(y',\cdot) \rangle \\ &= \mathbf{E}_{p,p} k(x,x') - 2\mathbf{E}_{p,q} k(x,y) + \mathbf{E}_{q,q} k(y,y')\end{aligned}$$

Maximum Mean Discrepancy Statistic

Goal: Estimate $D(p, q, \mathcal{F})$

$$\mathbf{E}_{p,p} k(x, x') - 2\mathbf{E}_{p,q} k(x, y) + \mathbf{E}_{q,q} k(y, y')$$

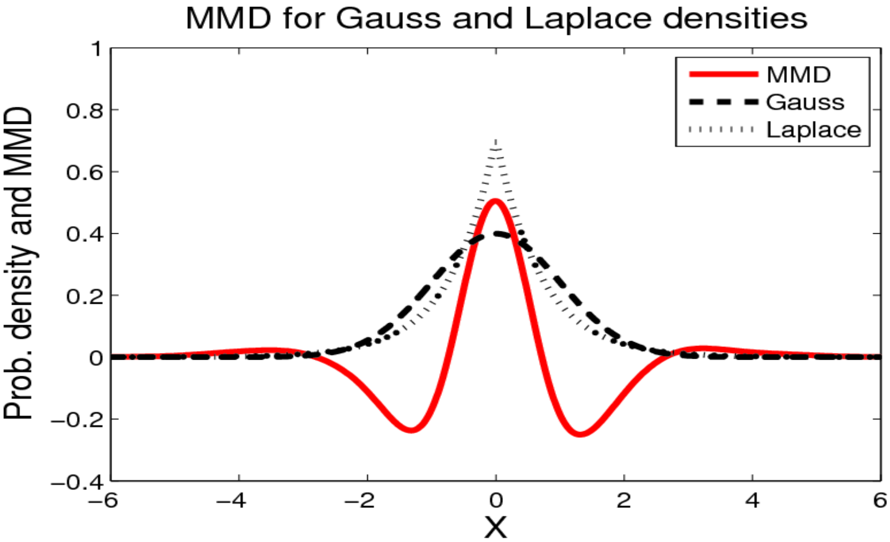
U-Statistic: Empirical estimate $D(X, Y, \mathcal{F})$

$$\frac{1}{m(m-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{=:h((x_i, y_i), (x_j, y_j))}$$

Theorem

$D(X, Y, \mathcal{F})$ is an unbiased estimator of $D(p, q, \mathcal{F})$.

Distinguishing Normal and Laplace



Uniform Convergence Bound

Theorem (Hoeffding, 1963)

For the kernel of a U-statistic $\kappa(x, x')$ *with* $|\kappa(x, x')| \leq r$ *we have*

$$\Pr\left\{\left|\mathbf{E}_p\left[\kappa(x, x')\right] - \frac{1}{m(m-1)}\sum_{i \neq j}\kappa(x_i, x_j)\right| > \epsilon\right\} \leq 2\exp\left(-\frac{m\epsilon^2}{r^2}\right)$$

Corollary (MMD Convergence)

$$\Pr\left\{|D(X, Y, \mathcal{F}) - D(p, q, \mathcal{F})| > \epsilon\right\} \leq 2\exp\left(-\frac{m\epsilon^2}{r^2}\right)$$

Consequences

- We have $O(\frac{1}{\sqrt{m}})$ *uniform convergence*, hence the estimator is *consistent*.
- We can use this as a test: solve the inequality for a given confidence level δ. Bounds can be **very** loose.

Uniform Convergence Bound

Theorem (Hoeffding, 1963)

For the kernel of a U-statistic $\kappa(x, x')$ *with* $|\kappa(x, x')| \leq r$ *we have*

$$\Pr\left\{\left|\mathbf{E}_p\left[\kappa(x, x')\right] - \frac{1}{m(m-1)}\sum_{i \neq j}\kappa(x_i, x_j)\right| > \epsilon\right\} \leq 2\exp\left(-\frac{m\epsilon^2}{r^2}\right)$$

Corollary (MMD Convergence)

$$\Pr\left\{\left|D(X, Y, \mathcal{F}) - D(p, q, \mathcal{F})\right| > \epsilon\right\} \leq 2\exp\left(-\frac{m\epsilon^2}{r^2}\right)$$

Consequences

- We have $O(\frac{1}{\sqrt{m}})$ *uniform convergence*, hence the estimator is *consistent*.
- We can use this as a test: solve the inequality for a given confidence level δ. Bounds can be **very** loose.

Uniform Convergence Bound

Theorem (Hoeffding, 1963)

For the kernel of a U-statistic $\kappa(x, x')$ *with* $|\kappa(x, x')| \leq r$ *we have*

$$\Pr\left\{\left|\mathbf{E}_p\left[\kappa(x, x')\right] - \frac{1}{m(m-1)} \sum_{i \neq j} \kappa(x_i, x_j)\right| > \epsilon\right\} \leq 2 \exp\left(-\frac{m\epsilon^2}{r^2}\right)$$

Corollary (MMD Convergence)

$$\Pr\left\{|D(X, Y, \mathcal{F}) - D(p, q, \mathcal{F})| > \epsilon\right\} \leq 2 \exp\left(-\frac{m\epsilon^2}{r^2}\right)$$

Consequences

- We have $O(\frac{1}{\sqrt{m}})$ *uniform convergence*, hence the estimator is *consistent*.
- We can use this as a test: solve the inequality for a given confidence level δ. Bounds can be **very** loose.

Asymptotic Bound

Idea

Use asymptotic normality of U-Statistic, estimate variance σ^2.

Theorem (Hoeffding, 1948)

$D(X, Y, \mathcal{F})$ asymptotically normal with variance $\frac{4\sigma^2}{m}$ and

$$\sigma^2 = \mathop{\mathbf{E}}_{x,y}\left[\left[\mathop{\mathbf{E}}_{x',y'} k((x,y),(x',y'))\right]^2\right] - \left[\mathop{\mathbf{E}}_{x,y,x',y'} k((x,y),(x',y'))\right]^2.$$

Test

- Estimate σ^2 from data.
- Reject hypothesis that $p = q$ if $D(X, Y, \mathcal{F}) > 2\alpha\sigma/\sqrt{m}$, where α is confidence threshold.
- Threshold is computed via $(2\pi)^{-\frac{1}{2}} \int_\alpha^\infty \exp(-x^2/2)dx = \delta$.

Asymptotic Bound

Idea

Use asymptotic normality of U-Statistic, estimate variance σ^2.

Theorem (Hoeffding, 1948)

$D(X, Y, \mathfrak{F})$ asymptotically normal with variance $\frac{4\sigma^2}{m}$ and

$$\sigma^2 = \mathop{\mathbf{E}}_{x,y}\left[\left[\mathop{\mathbf{E}}_{x',y'} k((x,y),(x',y'))\right]^2\right] - \left[\mathop{\mathbf{E}}_{x,y,x',y'} k((x,y),(x',y'))\right]^2.$$

Test

- Estimate σ^2 from data.
- Reject hypothesis that $p = q$ if $D(X, Y, \mathfrak{F}) > 2\alpha\sigma/\sqrt{m}$, where α is confidence threshold.
- Threshold is computed via $(2\pi)^{-\frac{1}{2}} \int_\alpha^\infty \exp(-x^2/2)dx = \delta$.

Asymptotic Bound

Idea

Use asymptotic normality of U-Statistic, estimate variance σ^2.

Theorem (Hoeffding, 1948)

$D(X, Y, \mathcal{F})$ asymptotically normal with variance $\frac{4\sigma^2}{m}$ and

$$\sigma^2 = \mathop{\mathbf{E}}_{x,y}\left[\left[\mathop{\mathbf{E}}_{x',y'} k((x,y),(x',y'))\right]^2\right] - \left[\mathop{\mathbf{E}}_{x,y,x',y'} k((x,y),(x',y'))\right]^2.$$

Test

- Estimate σ^2 from data.
- Reject hypothesis that $p = q$ if $D(X, Y, \mathcal{F}) > 2\alpha\sigma/\sqrt{m}$, where α is confidence threshold.
- Threshold is computed via $(2\pi)^{-\frac{1}{2}} \int_\alpha^\infty \exp(-x^2/2)dx = \delta$.

Outline

Application: Data Integration

Goal

- Data from various sources
- Check whether we can combine it

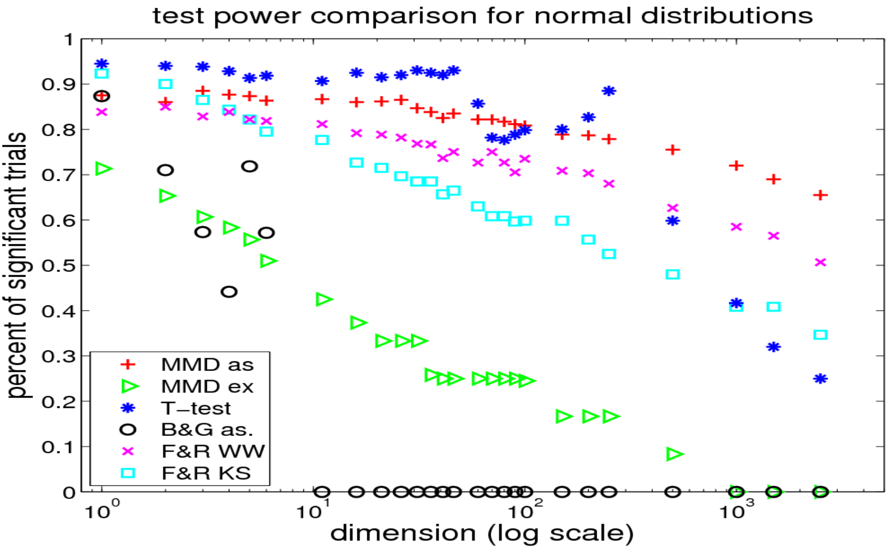
Comparison

- MMD using the uniform convergence bound
- MMD using the asymptotic expansion
- *t*-test
- Friedman-Rafsky Wolf test
- Friedman-Rafsky Smirnov test
- Hall-Tajvidi

Important Detail

Our test only needs a *double for loop* for implementation.
Other tests require spanning trees, matrix inversion, etc.

Toy Example: Normal Distributions



Microarray cross-platform comparability

Platforms	H_0	MMD	t-test	FR Wolf	FR Smirnov
Same	accepted	100	100	93	95
Same	rejected	0	0	7	5
Different	accepted	0	95	0	29
Different	rejected	100	5	100	71

- Cross-platform comparability tests on microarray level for cDNA and oligonucleotide platforms
  - repetitions: 100
  - sample size (each): 25
  - dimension of sample vectors: 2116

Cancer diagnosis

Health status	H_0	MMD	t-test	FR Wolf	FR Smirnov
Same	accepted	100	100	97	98
Same	rejected	0	0	3	2
Different	accepted	0	100	0	38
Different	rejected	100	0	100	62

- Comparing samples from normal and prostate tumor tissues. H_0 is hypothesis that $p = q$
  - repetitions 100
  - sample size (each) 25
  - dimension of sample vectors: 12,600

Tumor subtype tests

Subtype	H_0	MMD	t-test	FR Wolf	FR Smirnov
Same	accepted	100	100	95	96
Same	rejected	0	0	5	4
Different	accepted	0	100	0	22
Different	rejected	100	0	100	78

- Comparing samples from different and identical tumor subtypes of lymphoma. H_0 is hypothesis that $p = q$.
  - repetitions 100
  - sample size (each) 25
  - dimension of sample vectors: 2,118

Outline

Application: Attribute Matching

Goal

- Two datasets, find corresponding attributes.
- Use only distributions over random variables.
- Occurs when matching schemas between databases.

Examples

- Match different sets of dates
- Match names
- Can we merge the databases at all?

Approach

Use MMD to measure distance between distributions over different coordinates.

Application: Attribute Matching

Goal

- Two datasets, find corresponding attributes.
- Use only distributions over random variables.
- Occurs when matching schemas between databases.

Examples

- Match different sets of dates
- Match names
- Can we merge the databases at all?

Approach

Use MMD to measure distance between distributions over different coordinates.

Application: Attribute Matching

Goal

- Two datasets, find corresponding attributes.
- Use only distributions over random variables.
- Occurs when matching schemas between databases.

Examples

- Match different sets of dates
- Match names
- Can we merge the databases at all?

Approach

Use MMD to measure distance between distributions over different coordinates.

Application: Attribute Matching

Dataset	Attr.	MMD	MMD_u^2	t-test	FR Wolf	FR Smirnov	Hall	Biau
BIO	Same	**100.0**	99.3	95.2	90.3	95.8	95.3	99.3
	Different	20.0	19.8	36.2	**17.2**	18.6	17.9	42.1
FOREST	Same	**100.0**	**100.0**	97.4	94.6	99.8	95.5	**100.0**
	Different	8.1	1.5	0.2	3.8	**0.0**	50.1	**0.0**
CNUM	Same	**100.00**	99.29	95.00	98.14	99.00	84.86	99.43
	Different	17.58	**5.37**	16.82	24.63	14.07	81.65	48.48
FOREST10D	Same	**100.0**	98.0	**100.0**	93.5	96.5	97.0	**100.0**
	Different	100.0	3.0	**0.0**	**0.0**	1.0	72.0	100.0

Linear Assignment Problem

Goal

Find good assignment for all pairs of coordinates (i, j).

$$\underset{\pi}{\text{maximize}} \sum_{i=1}^{m} C_{i\pi(i)} \text{ where } C_{ij} = D(X_i, X_j', \mathcal{F})$$

Optimize over the space of all permutation matrices π.

Linear Programming Relaxation

$$\underset{\pi}{\text{maximize}} \ \operatorname{tr} C^\top \pi$$

$$\text{subject to } \sum_i \pi_{ij} = 1 \text{ and } \sum_j \pi_{ij} = 1 \text{ and } \pi_{ij} \geq 0 \text{ and } \pi_{ij} \in \{0, 1\}$$

Integrality constraint can be dropped, as the remainder of the matrix is unimodular.

Hungarian Marriage (Kuhn, Munkres, 1953)

Solve in cubic time.

Linear Assignment Problem

Goal

Find good assignment for all pairs of coordinates (i, j).

$$\underset{\pi}{\text{maximize}} \sum_{i=1}^{m} C_{i\pi(i)} \text{ where } C_{ij} = D(X_i, X'_j, \mathcal{F})$$

Optimize over the space of all permutation matrices π.

Linear Programming Relaxation

$$\underset{\pi}{\text{maximize}} \ \operatorname{tr} C^\top \pi$$

$$\text{subject to } \sum_i \pi_{ij} = 1 \text{ and } \sum_j \pi_{ij} = 1 \text{ and } \pi_{ij} \geq 0 \text{ and } \pi_{ij} \in \{0, 1\}$$

Integrality constraint can be dropped, as the remainder of the matrix is unimodular.

Hungarian Marriage (Kuhn, Munkres, 1953)

Solve in cubic time.

Linear Assignment Problem

Goal

Find good assignment for all pairs of coordinates (i, j).

$$\underset{\pi}{\text{maximize}} \sum_{i=1}^{m} C_{i\pi(i)} \text{ where } C_{ij} = D(X_i, X'_j, \mathcal{F})$$

Optimize over the space of all permutation matrices π.

Linear Programming Relaxation

$$\underset{\pi}{\text{maximize}} \ \operatorname{tr} C^\top \pi$$

$$\text{subject to } \sum_i \pi_{ij} = 1 \text{ and } \sum_j \pi_{ij} = 1 \text{ and } \pi_{ij} \geq 0 \text{ and } \pi_{ij} \in \{0, 1\}$$

Integrality constraint can be dropped, as the remainder of the matrix is unimodular.

Hungarian Marriage (Kuhn, Munkres, 1953)

Solve in cubic time.

Schema Matching with Linear Assignment

Key Idea

Use $D(X_i, X'_j, \mathcal{F}) = C_{ij}$ as compatibility criterion.

Results

Dataset	Data	d	m	rept.	% correct
BIO	uni	6	377	100	92.0
CNUM	uni	14	386	100	99.8
FOREST	uni	10	538	100	100.0
FOREST10D	multi	2	1000	100	100.0
ENYZME	struct	6	50	50	100.0
PROTEINS	struct	2	200	50	100.0

Outline

Sample Bias Correction

The Problem

- Training data X, Y is drawn iid from $\Pr(x, y)$.
- Test data X', Y' is drawn iid from $\Pr'(x', y')$.
- Simplifying assumption: only $\Pr(x)$ and $\Pr'(x')$ differ. Conditional distributions $\Pr(y|x)$ are the same.

Applications

- In medical diagnosis (e.g. cancer detection from microarrays) we usually have very different training and test sets.
- Active learning
- Experimental design
- Brain computer interfaces (drifting distributions)
- Adapting to new users

Sample Bias Correction

The Problem

- Training data X, Y is drawn iid from $\Pr(x, y)$.
- Test data X', Y' is drawn iid from $\Pr'(x', y')$.
- Simplifying assumption: only $\Pr(x)$ and $\Pr'(x')$ differ. Conditional distributions $\Pr(y|x)$ are the same.

Applications

- In medical diagnosis (e.g. cancer detection from microarrays) we usually have very different training and test sets.
- Active learning
- Experimental design
- Brain computer interfaces (drifting distributions)
- Adapting to new users

Importance Sampling

Swapping Distributions

Assume that we are allowed to draw from p but want to draw from q:

$$\mathbf{E}_q\left[f(x)\right] = \int \underbrace{\frac{q(x)}{p(x)}}_{\beta(x)} f(x) dp(x) = \mathbf{E}_p\left[\frac{q(x)}{p(x)} f(x)\right]$$

Reweighted Risk

Minimize reweighted empirical risk plus regularizer, as in SVM, regression, GP classification.

$$\frac{1}{m}\sum_{i=1}^{m} \beta(x_i) l(x_i, y_i, \theta) + \lambda\Omega[\theta]$$

Problem We need to know $\beta(x)$.

Problem We are ignoring the fact that we know the test set.

Importance Sampling

Swapping Distributions

Assume that we are allowed to draw from p but want to draw from q:

$$\mathbf{E}_q\left[f(x)\right] = \int \underbrace{\frac{q(x)}{p(x)}}_{\beta(x)} f(x) dp(x) = \mathbf{E}_p\left[\frac{q(x)}{p(x)} f(x)\right]$$

Reweighted Risk

Minimize reweighted empirical risk plus regularizer, as in SVM, regression, GP classification.

$$\frac{1}{m}\sum_{i=1}^{m} \beta(x_i) l(x_i, y_i, \theta) + \lambda\Omega[\theta]$$

Problem We need to know $\beta(x)$.

Problem We are ignoring the fact that we know the test set.

Importance Sampling

Swapping Distributions

Assume that we are allowed to draw from p but want to draw from q:

$$\mathbf{E}_q\left[f(x)\right] = \int \underbrace{\frac{q(x)}{p(x)}}_{\beta(x)} f(x) dp(x) = \mathbf{E}_p\left[\frac{q(x)}{p(x)} f(x)\right]$$

Reweighted Risk

Minimize reweighted empirical risk plus regularizer, as in SVM, regression, GP classification.

$$\frac{1}{m}\sum_{i=1}^{m} \beta(x_i) l(x_i, y_i, \theta) + \lambda\Omega[\theta]$$

Problem We need to know $\beta(x)$.

Problem We are ignoring the fact that we know the test set.

Importance Sampling

Swapping Distributions

Assume that we are allowed to draw from p but want to draw from q:

$$\mathbf{E}_q[f(x)] = \int \underbrace{\frac{q(x)}{p(x)}}_{\beta(x)} f(x) dp(x) = \mathbf{E}_p\left[\frac{q(x)}{p(x)} f(x)\right]$$

Reweighted Risk

Minimize reweighted empirical risk plus regularizer, as in SVM, regression, GP classification.

$$\frac{1}{m}\sum_{i=1}^{m} \beta(x_i) l(x_i, y_i, \theta) + \lambda\Omega[\theta]$$

Problem We need to know $\beta(x)$.

Problem We are ignoring the fact that we know the test set.

Reweighting Means

Theorem

The optimization problem

$$\underset{\beta(x)\geq 0}{\text{minimize}} \; \left\| \mathbf{E}_q\left[k(x,\cdot)\right] - \mathbf{E}_p\left[\beta(x)k(x,\cdot)\right]\right\|^2$$

$$\text{subject to } \mathbf{E}_p\left[\beta(x)\right] = 1$$

is convex and its unique solution is $\beta(x) = q(x)/p(x)$.

Proof.

1. The problem is obviously convex: convex objective and linear constraints. Moreover, it is bounded from below by 0.
2. Hence it has a unique minimum.
3. The "guess" $\beta(x) = q(x)/p(x)$ achieves the minimum value. □

Reweighting Means

Theorem

The optimization problem

$$\underset{\beta(x)\geq 0}{\textit{minimize}} \; \left\| \mathbf{E}_q\left[k(x,\cdot)\right] - \mathbf{E}_p\left[\beta(x)k(x,\cdot)\right] \right\|^2$$

$$\textit{subject to } \mathbf{E}_p\left[\beta(x)\right] = 1$$

is convex and its unique solution is $\beta(x) = q(x)/p(x)$.

Proof.

1. The problem is obviously convex: convex objective and linear constraints. Moreover, it is bounded from below by 0.
2. Hence it has a unique minimum.
3. The "guess" $\beta(x) = q(x)/p(x)$ achieves the minimum value. $\square$

NATIONAL ICT AUSTRALIA LIMITED

Empirical Version

- Re-weight empirical mean in feature space

$$\mu[X, \beta] := \frac{1}{m} \sum_{i=1}^{m} \beta_i k(x_i, \cdot)$$

such that it is close to the mean on test set

$$\mu[X'] := \frac{1}{m'} \sum_{i=1}^{m'} k(x_i', \cdot)$$

- Ensure that β_i is proper reweighting.

Optimization Problem

Quadratic Program

$$\underset{\beta}{\text{minimize}} \left\| \frac{1}{m} \sum_{i=1}^{m} \beta_i k(x_i, \cdot) - \frac{1}{m'} \sum_{i=1}^{m'} k(x'_i, \cdot) \right\|^2$$

$$\text{subject to } 0 \leq \beta_i \leq B \text{ and } \sum_{i=1}^{m} \beta_i = m$$

- Upper bound on β_i for regularization
- Summation constraint for reweighted distribution
- Standard solvers available

Consistency

Theorem

- *The reweighted set of observations will behave like one drawn from q with effective sample size $m^2 / \|\beta\|^2$.*
- *The bias of the estimator is proportional to the square root of the value of the objective function.*

Proof.

1. Show that for smooth functions expected loss is close. This only requires that both feature map means are close.
2. Show that expected loss is close to empirical loss (in a transduction style, i.e. conditioned on X and X'.

Consistency

Theorem

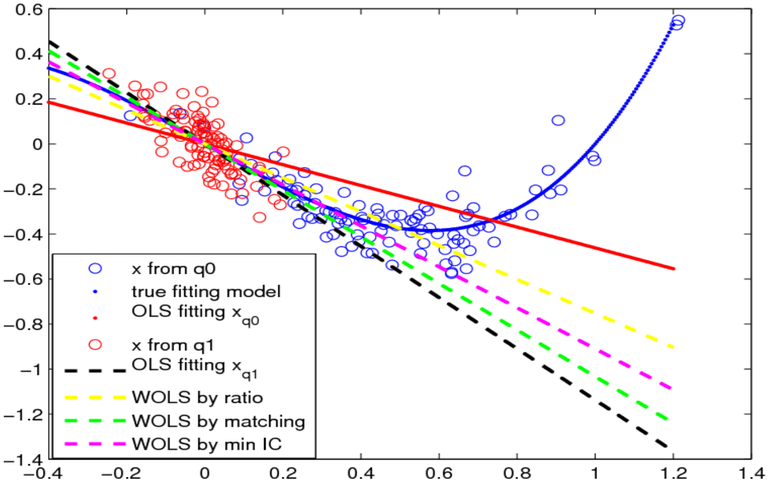
- *The reweighted set of observations will behave like one drawn from q with effective sample size* $m^2 / \|\beta\|^2$*.*
- *The bias of the estimator is proportional to the square root of the value of the objective function.*

Proof.

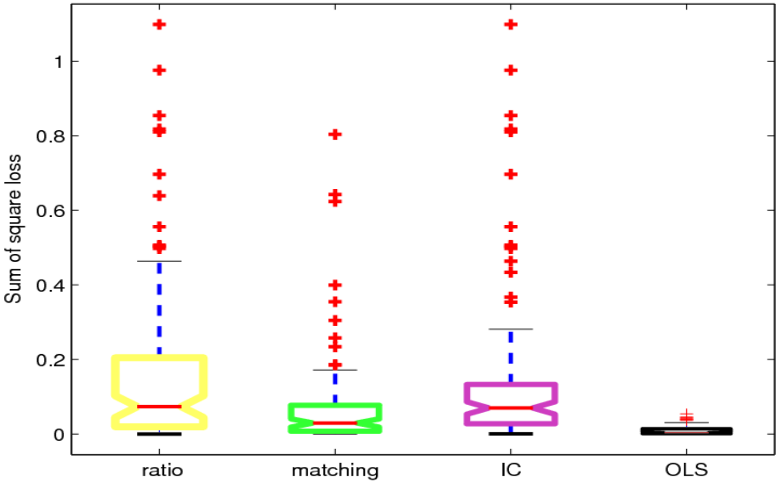
1. Show that for smooth functions expected loss is close. This only requires that both feature map means are close.
2. Show that expected loss is close to empirical loss (in a transduction style, i.e. conditioned on X and X'.

□

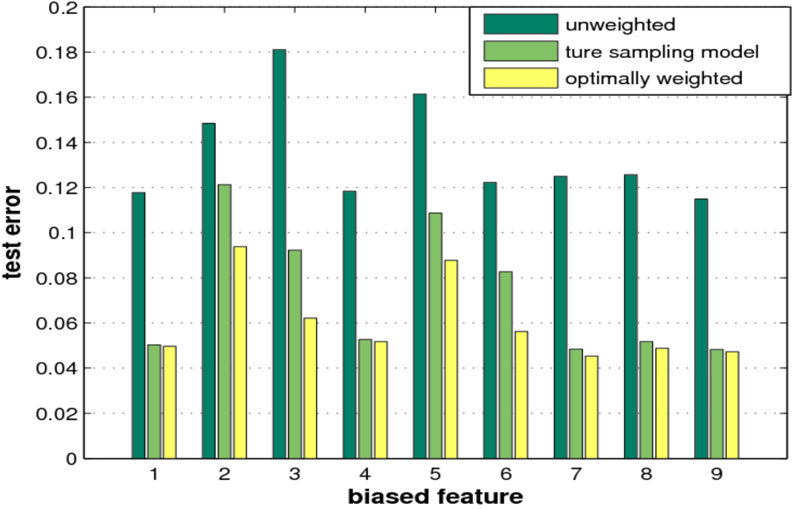
Regression Toy Example



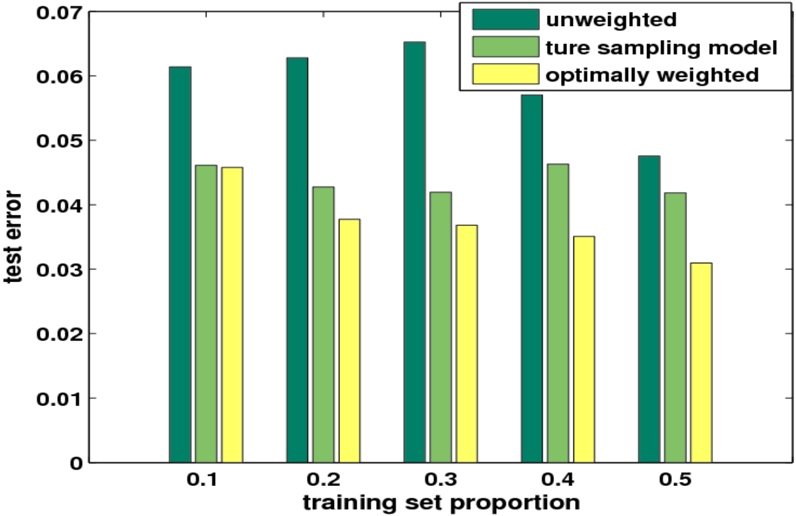
Regression Toy Example



Breast Cancer - Bias on features



Breast Cancer - Bias on labels



More Experiments

DataSet	n_{tr}	selected	n_{tst}	unweighted	NMSE / Test err. import. sampling	KMM
1. Abalone*	2000	853	2177	1.00 ± 0.08	1.1 ± 0.2	$\mathbf{0.6 \pm 0.1}$
2. CA Housing*	16512	3470	4128	2.29 ± 0.01	1.72 ± 0.04	$\mathbf{1.24 \pm 0.09}$
3. Delta Ailerons(1)*	4000	1678	3129	0.51 ± 0.01	0.51 ± 0.01	$\mathbf{0.401 \pm 0.007}$
4. Ailerons*	7154	925	6596	1.50 ± 0.06	$\mathbf{0.7 \pm 0.1}$	1.2 ± 0.2
5. haberman(1)	150	52	156	0.50 ± 0.09	0.37 ± 0.03	$\mathbf{0.30 \pm 0.05}$
6. USPS(6vs8)(1)	500	260	1042	0.13 ± 0.18	0.1 ± 0.2	$\mathbf{0.1 \pm 0.1}$
7. USPS(3vs9)(1)	500	252	1145	0.016 ± 0.006	$\mathbf{0.012 \pm 0.005}$	0.013 ± 0.005
8. Bank8FM*	4500	654	3692	0.5 ± 0.1	$\mathbf{0.45 \pm 0.06}$	0.47 ± 0.05
9. Bank32nh*	4500	740	3692	23 ± 4.0	$\mathbf{19 \pm 2}$	$\mathbf{19 \pm 2}$
10. cpu-act*	4000	1462	4192	10 ± 1	4.0 ± 0.2	$\mathbf{1.9 \pm 0.2}$
11. cpu-small*	4000	1488	4192	9 ± 2	4.0 ± 0.2	$\mathbf{2.0 \pm 0.5}$
12. Delta Ailerons(2)*	4000	634	3129	2 ± 2	$\mathbf{1.5 \pm 1.5}$	1.7 ± 0.9
13. Boston house*	300	108	206	0.8 ± 0.2	$\mathbf{0.74 \pm 0.09}$	0.76 ± 0.07
14. kin8nm*	5000	428	3192	0.85 ± 0.2	$\mathbf{0.81 \pm 0.1}$	0.81 ± 0.2
15. puma8nh*	4499	823	3693	1.1 ± 0.1	$\mathbf{0.77 \pm 0.05}$	0.83 ± 0.03
16. haberman(2)	150	90	156	0.27 ± 0.01	0.39 ± 0.04	$\mathbf{0.25 \pm 0.2}$
17. USPS(6vs8) (2)	500	156	1042	0.23 ± 0.2	0.23 ± 0.2	$\mathbf{0.16 \pm 0.08}$
18. USPS(6vs8) (3)	500	104	1042	0.54 ± 0.0002	0.5 ± 0.2	$\mathbf{0.16 \pm 0.04}$
19. USPS(3vs9)(2)	500	252	1145	0.46 ± 0.09	0.5 ± 0.2	$\mathbf{0.2 \pm 0.1}$
20. Breast Cancer	280	96	419	0.05 ± 0.01	0.036 ± 0.005	$\mathbf{0.033 \pm 0.004}$
21. Indias diabets	200	97	568	0.32 ± 0.02	$\mathbf{0.30 \pm 0.02}$	$\mathbf{0.30 \pm 0.02}$
22. ionosphere	150	64	201	0.32 ± 0.06	0.31 ± 0.07	$\mathbf{0.28 \pm 0.06}$
23. German credit	400	214	600	0.283 ± 0.004	0.282 ± 0.004	$\mathbf{0.280 \pm 0.004}$

Summary

1. **Two Sample Problem**
   - Direct Solution
   - Kolmogorov Smirnov Test
   - Reproducing Kernel Hilbert Spaces
   - Test Statistics
2. **Data Integration**
   - Problem Definition
   - Examples
3. **Attribute Matching**
   - Basic Problem
   - Linear Assignment Problem
4. **Sample Bias Correction**
   - Sample Reweighting
   - Quadratic Program and Consistency
   - Experiments