Kernel Methods

Lecture 3: Inference and Convex Duality
Thanks to Yasemin Altun, Markus Hegland

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

Inverse Problems

Observations

- Data $x_1, \ldots, x_m$, drawn from some distribution $p(x)$.
- Measurements $y_1, \ldots, y_m$ observed at $x_1, \ldots, x_m$.
- Indirect measurements $y_1, \ldots, y_m$ generated by some measurement process A.

Formal Definition

Solve the problem $Ax = b$ for unknown x.

Ill Posed Problem

We do not have enough data to find x exactly — A does not have full rank.

Solution

Solve a regularized risk minimization problem.

$$\text{minimize } f(x) \text{ subject to } \|Ax - b\| = 0$$

Example

Density Estimation

We want to have a density matching empirical means

$$\mathbf{E}[x] = \frac{1}{m}\sum_{i=1}^{m} x_i \text{ and } \mathbf{E}\left[x^2\right] = \frac{1}{m}\sum_{i=1}^{m} x_i^2$$

Ill Posed Problem

- Many distributions possible, e.g. $p(x) = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}(x)$.
- We need a regularity condition

Regularizers

- Small squared norm of the density $\int p^2(x)dx$.
- Density should be smooth, i.e. small $\int |\partial_x p(x)|^2 \, dx$.
- Non-informative density, i.e. small entropy

$$\int -p(x)\log p(x)dx.$$

Maximum Entropy Principle

Motivation

Find least informative consistent distribution.

Moment Matching

Given $\phi(x)$ find p such that $\mathbf{E}[\phi(x)] = \hat{\mu} := \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$.

Theorem (MaxEnt is dual to Maximum Likelihood)

$$\text{minimize} \int -p(x)\log p(x)dx \text{ subject to } \mathbf{E}[\phi(x)] = \hat{\mu}$$

has as its dual the maximum likelihood problem

$$\text{minimize } g(\theta) - \langle \hat{\mu}, \theta \rangle \text{ where } g(\theta) = \int \exp(\langle \phi(x), \theta \rangle)\, dx.$$

Proof

Lagrangian

We need to ensure that p is nonnegative and is normalized:

$$L(p,\theta,\Lambda,\eta) = \int -p(x)\log p(x)dx + \lambda^\top\left[\hat{\mu} - \int \phi(x)p(x)dx\right]$$
$$+ \Lambda\left[1 - \int p(x)dx\right] + \int \eta(x)p(x)dx$$

Variational Derivative

Informally, we can pull the derivative into the integral:

$$\partial_p L(p,\theta,g,\eta) = -\log p(x) + 1 + \theta^\top\phi(x) + \Lambda + \eta(x) = 0$$

Solution

$$p(x) = \exp\left(\theta^\top\phi(x) + \underbrace{\Lambda + 1}_{:=-g(\theta)} + \underbrace{\eta(x)}_{=0}\right)$$

Proof (Part II)

Wolfe's Dual

Plugging the expansion $p(x) = (\langle \phi(x), \theta \rangle - g(\theta))$ into the Lagrangian yields:

$$\text{maximize } \lambda^\top \hat{\mu} - g(\theta) \text{ where } g(x) = \log \int \exp\left(\lambda^\top \phi(x)\right) dx.$$

Maximum Likelihood

Expanding the dual objective by m and using $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$ proves the claim.

Caveat

We ignored **feasibility** and **constraint qualification** of the problem.

NATIONAL ICT AUSTRALIA LIMITED

Approximate Moment Matching

Exact Moment Matching

This requires that the distribution has exactly the same moments as the empirical mean.

Example: Estimating a Normal Distribution

- Normal distribution with 0 mean and variance 1.
- Empirical average of x is 0.03, that of x^2 is 1.07.
- Clearly exact moment matching is unrealistic!

Solution

Maximum entropy under approximate moment matching:

$$\text{minimize} \int -p(x) \log p(x) dx \text{ subject to } \left\| \mathbf{E}\left[\phi(x)\right] - \hat{\mu} \right\| \leq \epsilon$$

Previous Work

General Problem

$$\text{minimize } f(x) \text{ subject to } \|Ax - b\| \leq \epsilon$$

AdaBoost (Lafferty 1999, Kivinen etc 1999, Colins etc 2000)

- $f(x)$ is the Bregmann divergence corresponding to the unnormalized entropy.
- $\epsilon = 0$, $b = 0$ and A takes care of deviations from the empirical averages (more later).

Regularized MaxEnt (Dudik et al., 2004, 2006)

- $f(x)$ is the (normalized) entropy
- b is empirical average of moments and A is expectation operator. $\|\cdot\|$ are ℓ_1 and ℓ_2 norms.

Regularization Theory (Arsenin and Tikhonov, 1977)

- $f(x)$ is $\|x\|^2$ squared ℓ_2 norm.
- General ill-posed problem $Ax = b$.

Questions

General Case
Unified treatment of approximate solutions of ill-posed problems.

Algorithm
Efficient algorithm to solve all the problems.

Feasibility
When is the problem feasible, bounded, etc.? When can we compute the dual?

Interpretation
What is the meaning of the dual problem?

Fenchel Duality

Definition (Convex Conjugate)

Denote by $f : \mathcal{X} \to \mathbb{R}$ a convex function on some convex domain $\mathcal{X}$ of a Banach space $\mathcal{B}$. Then the dual $f^* : \mathcal{B}^* \to \mathbb{R}$ is defined as

$$f^*(x^*) := \sup_{x \in \mathcal{X}} \langle x, x^* \rangle - f(x).$$

Properties

Self Duality $f^{**} = f$

Linear Offset $\{f(x) + \langle a, x \rangle\}^* = f^*(x^* - a)$

Linear Functions

$f(x) = \langle a, x \rangle$ and $\mathcal{X} = U_{\mathcal{B}}(1)$ implies $f^*(x^*) = \|x^* - a\|$.

Key Theorem

Theorem (Fenchel's Duality with Constraints)

$$t := \inf_{x \in \mathcal{X}} \{ f(x) \text{ subject to } \|Ax - b\|_{\mathcal{B}} \}$$

$$d := \sup_{x^* \in \mathcal{B}^*} \{ -f^*(A^* x^*) + \langle b, x^* \rangle - \epsilon \|x^*\|_{\mathcal{B}^*} \}$$

If $\mathrm{core}(A \,\mathrm{dom}\, f) \cap (b + \epsilon \,\mathrm{int}(B)) \neq \emptyset$ *then* $t = d$.

- Note that $s \in \mathrm{core}(S)$ if $\bigcup_{\lambda > 0} \lambda(S - s) \subseteq \mathcal{X}$ and $S \in \mathcal{X}$.
- This is the price we pay for infinite dimensionality.
- This allows us to rewrite optimization problems in the dual domain.

Application: Csiszar Divergence

Divergence

Denote by q a reference density and let h be convex.

$$f(p) := \int q(t) h\left(\frac{p(t)}{q(t)}\right) dt$$

Special cases are Tsallis, Burg, Amari, and KL divergence.

Primal Problem

$$\underset{p}{\text{minimize}}\, f(p) \text{ subject to } \left\|\mathbf{E}\left[\phi(x)\right] - \hat{\mu}\right\|_{\mathcal{B}} \leq \epsilon \text{ and } \int dp = 1$$

Dual Problem

$$\underset{\theta}{\text{maximize}} - \int q(t) h^{*}\left(\langle \theta, \phi(t)\rangle - \Lambda\right) dt + \langle \theta, \hat{\mu}\rangle - \Lambda - \epsilon \left\|\theta\right\|_{\mathcal{B}^*}$$

Density is given by $p(t) = q(t)(h^{*})'\left(\langle \theta, \phi(t)\rangle - \Lambda\right)$.

Application: KL-Divergence

Divergence

$h(\xi) = \xi \log \xi$ yields Kullback-Leibler divergence.

Dual Problem

$$\underset{\theta}{\text{maximize}} - \log \int q(t) \exp\left(\langle \theta, \phi(t) \rangle\right) dt + \langle \theta, \hat{\mu} \rangle - \epsilon \left\|\theta\right\|_{\mathcal{B}^*} + e^{-1}$$

Density is given by $p(t) = q(t) \exp\left(\langle \theta, \phi(t) \rangle - g(\theta)\right)$.

Examples

- For $\mathcal{B} = \ell_\infty$ we get ℓ_1 penalization (Dudik et al. 2004).
- For $\mathcal{B} = \mathcal{H}$ we get a kernel method (Nemermann and Bialek, 1998)

Application: Conditional Models

AdaBoost (Collins et al., 2000)

- f is sum over unnormalized entropies for $p(y|x_i)$.
- A is sum over evaluations of features at locations (x_i, y_i)
- $\epsilon = 0$

Gaussian Process Classification

- f is sum over normalized entropies for $p(y_i|x_i)$.
- A is sum over evaluations of features at locations (x_i, y_i)
- $\mathcal{B}$ is a Hilbert Space

Gaussian Process Regression

Same as classification, only different sufficient statistics.

Conditional Random Fields

Sufficient statistics $\phi(x, y)$ decomposes into cliques.

Concentration of Empirical Means

Problem

We need to determine ϵ for the constraint

$$\left\| \mathbf{E}\left[\phi(x)\right] - \frac{1}{m}\sum_{i=1}^{m}\phi(x_i) \right\| \leq \epsilon$$

Theorem (Uniform Convergence to Empirical Means)

With probability $\delta \leq 1 - \exp\left(-\frac{\epsilon^2 m}{R^2}\right)$

$$\left\| \mathbf{E}\left[\phi(x)\right] - \hat{\mu} \right\| \leq 2R_m(\mathfrak{F}, p) + \epsilon$$

R_m is Rademacher average. $\mathfrak{F}$ is the class of linear functions of bounded norm.

Advantage

Principled regularization scheme $\epsilon = O(m^{-1/2})$.

Risk Bounds

Loss

$$L(\theta, \mu) := f^*(\langle \theta, \phi(\cdot) \rangle) - \langle \mu, \theta \rangle + \epsilon \, \|\theta\|_{\mathcal{B}^*}^k$$

True Statistics Let μ^* be the true mean.

Theorem

With probability $\delta \leq 1 - \exp\left(-\frac{\epsilon^2 m}{R^2}\right)$

$$L(\theta^*, \mu^*) - L(\theta, \mu) \leq \|\theta^*\| \left[2R_m(\mathcal{F}, p) + \epsilon\right]$$

Proof.

Use the relation $L(\theta, \mu) - L(\theta^*, \mu^*) \leq \langle \theta, \mu^* - \mu \rangle$ and the concentration of empirical means. $\square$

Risk Bounds

We want to bound the deviation between the loss at the actual solution θ and the optimal solution θ^*.

Theorem

With probability $\delta \leq 1 - \exp\left(-\frac{\epsilon^2 m}{R^2}\right)$

$$L(\theta, \mu^*) - L(\theta^*, \mu^*) \leq 2\left[\frac{C}{k\epsilon}\right]^{1/k-1}\left[2R_m(\mathfrak{F}, p) + \epsilon\right]$$

In the case of an RKHS we can get slightly tighter bounds instead of using Rademacher averages.

Optimization

We can use Zhang's algorithm (2003) for minimization.

1: **input:** sample of size m, statistics ϕ, base function class $\mathcal{B}^*_{\mathrm{base}}$, approximation ϵ, number of iterations K, and radius of the space of solutions R

2: Set $\theta = 0$.

3: **for** $k = 1, \ldots, K$ **do**

4: Find $(\hat{\imath}, \hat{\lambda})$ such that for $e_i \in \mathcal{B}^*_{\mathrm{base}}$ and $\lambda \in [0, 1]$ the following is approximately minimized:

$$L((1 - \lambda)\phi + R\lambda e_i, b)$$

5: Update $\phi \leftarrow (1 - \hat{\lambda})\phi + R\hat{\lambda}e_{\hat{\imath}}$

6: **end for**

This gives us $O(1/K)$ rate of convergence.

Shameless Plugs

Looking for a job ... talk to me!

- Alex.Smola@nicta.com.au (http://www.nicta.com.au)

Positions

- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels

- **http://sml.nicta.com.au**
- **http://www.kernel-machines.org**
- **http://www.learning-with-kernels.org**
  Schölkopf and Smola: Learning with Kernels