

Learning in Structured Domains

Candidacy exam

Risi Kondor

The Formal Framework

Learning from labeled examples: supervised learning

Known spaces \mathcal{X} and \mathcal{Y} ;

Unknown distribution P on $\mathcal{X} \times \mathcal{Y}$;

Training examples: $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ sampled from P ;

Goal is to learn $f: \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes $\mathbb{E}[L(f(x), y)]$ for some pre-defined loss function L .

Special cases (examples):

Classification	$\mathcal{Y} = \{-1, +1\}$	$L = (1 - f(x)y) / 2$
Regression	$\mathcal{Y} = \mathbb{R}$	$L = (f(x) - y)^2$

Algorithm selects some \hat{f} from some class of functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Empirical vs. true errors:

$$R_{\text{emp}}[f] = \frac{1}{m} \sum_{i=1}^m L(f(x_i), y_i). \quad \longleftrightarrow \quad R[f] = \mathbb{E}_P [L(f(x), y)]$$

R_{emp} vs. R and Generalization Bounds

Function returned by algorithm is not independent of sample and likely to be close to worst in \mathcal{F} , therefore interested in

$$\sup_{f \in \mathcal{F}} R[f] - R_{\text{emp}}[f].$$

$R_{\text{emp}}[g]$ is a random variable, therefore can only bound it probabilistically:

$$\sup_P \mathbb{P} \left[\left| \sup_{f \in \mathcal{F}} R[f] - R_{\text{emp}}[f] \right| \geq \epsilon \right] < \delta. \quad (1)$$

Introducing $\mathcal{F}_L = \{ L \circ f \mid f \in \mathcal{F} \}$, (1) becomes a statement about the deviations of the **empirical process**

$$\sup_{\hat{f}_L \in \mathcal{F}_L} (P f_L - P_n f_L).$$

We have a **uniform Glivenko-Cantelli class** when δ goes to zero as $m \rightarrow \infty$.

Empirical Risk Minimization

Algorithm selects f by minimizing some (possibly modified version) of R_{emp} . To guard against overfitting

1. restrict \mathcal{G} or
2. add complexity penalty term.

Regularized Risk Minimization:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \underbrace{R_{\text{emp}}[f] + \Omega[f]}_{R_{\text{reg}}[f]}.$$

Ill-posed problems, inverse problems, etc.

Hilbert Space Methods

[Schölkopf 2002] [Girosi 1993] [Smola 1998]

Hilbert space methods

Start with a regularized risk functional of the form

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^m L(f(x_i), y_i) + \frac{\lambda}{2} \|f\|^2$$

where $\|f\|^2 = \langle f, f \rangle_{\mathcal{F}}$ and \mathcal{F} is the RKHS induced by some positive definite kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Letting $k_x = k(x, \cdot)$, the RKHS is the closure of

$$\left\{ \sum_{i=1}^n \alpha_i k_{x_i} \mid n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

with respect to the inner product generated by $\langle k_x, k_{x'} \rangle = k(x, x')$. One consequence is the reproducing property:

$$\langle f, k_x \rangle = f(x).$$

Hilbert Space Methods

By reproducing property $R_{\text{reg}}[f]$ becomes

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^m L(\langle f, k_{x_i} \rangle, y_i) + \frac{\lambda}{2} \|f\|^2 \quad (2)$$

reducing problem to linear algebra, a quadratic problem, or something similar.

Representer theorem: solution to (2) is of form

$$f = \sum_{i=1}^m \alpha_i k_{x_i}.$$

Algorithm is determined by form of L and regularization scheme is determined by the kernel.

Regularization and kernels

Define operator $K : L_2(\mathcal{X}) \rightarrow L_2(\mathcal{X})$ as

$$(Kg)(x) = \int_{\mathcal{X}} k(x, x')g(x') dx.$$

For $f = Kg \in \mathcal{F}$, norm becomes

$$\langle f, f \rangle = \int_{\mathcal{X}} \int_{\mathcal{X}} g(x)g(x')k(x, x') dx dx' = \langle g, Kg \rangle_{L_2} = \langle f, K^{-1}f \rangle_{L_2}$$

Another way to approach this is from a **regularization network**

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^m L(\langle f, k_{x_i} \rangle, y_i) + \frac{\lambda}{2} \| Pf \|_{L_2}^2$$

for some regularization operator P . The kernel then becomes the **Green's function** of $P^\dagger P$:

$$P^\dagger P k(x, \cdot) = \delta_x.$$

Regularization and kernels

By Bochner's theorem, for translation invariant kernels ($k(x, x') = k(x - x')$), Fourier transform $\tilde{k}(\omega)$ is pointwise positive.

For Gaussian RBF kernel $k(x) = e^{-x^2/(2\sigma^2)}$ and $\tilde{k}(\omega) = e^{-\omega^2\sigma^2/2}$, so regularization term is

$$\langle f, f \rangle = \int e^{\omega^2\sigma^2/2} |\tilde{f}(\omega)|^2 d\omega = \sum_{m=0}^{\infty} \int \frac{\sigma^{2m}}{2^m m!} \| (O^m f)(x) \|_{L_2} dx$$

where $O^{2m} = \Delta^m$ and $O^{2m+1} = \nabla \Delta^m$. This is a natural notion of smoothness for functions.

A more exotic example are B_q spline kernels [Vapnik 1997]

$$k(x) = \prod_{i=1}^n B_q(x_i) \quad B_q = \otimes^{q+1} \mathbf{1}_{[-0.5, 0.5]} \quad \langle f, f \rangle = \int \left(\frac{\sin(\omega/2)}{\omega/2} \right)^{-q-1} |\tilde{f}(\omega)|^2 d\omega.$$

Gaussian Processes/Ridge Regression

Definition: collection of random variables $\{t_x\}$ indexed by $x \in \mathcal{X}$ such that any finite subset is jointly Gaussian distributed. Defined by mean $\mu(x) = \mathbb{E}[t_x]$ and covariance $\text{Cov}(t_x, t_{x'})$.

Assume $\mu = 0$ and $y_i \sim \mathcal{N}(0, \sigma_n^2)$. Then MAP estimate is minimizer of

$$R_{\text{reg}}[f] = \frac{1}{\sigma_n^2} \sum_{i=1}^m (\langle f, k_{x_i} \rangle, y_i)^2 + \|f\|^2$$

with kernel $k(x, x') = \text{Cov}(t_x, t_{x'})$. Solution is simply $f_{\text{MAP}}(x) = \vec{k} (K + \sigma^2 I)^{-1} \vec{y}^\top$ where $\vec{k} = (k(x, x_1), \dots, k(x, x_m))$ and $[K]_{i,j} = k(x_i, x_j)$ [Mackay 1997].

Support Vector Machines

Define feature map $\Phi : x \mapsto k_x$. Finds maximum margin separating hyperplane between images in RKHS... In feature space $f(x) = \text{sgn}^+(b + w \cdot x)$ where w is the solution of

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y_i (w \cdot x_i + b) \geq 1.$$

Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i (y_i (w \cdot x_i + b) - 1).$$

gives $\sum_{i=1}^m \alpha_i y_i = 0$ and $w = \sum_{i=1}^m \alpha_i y_i x_i$ leading to the dual problem

$$\max_{\alpha} \left[\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \right] \quad \text{s.t.} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^m \alpha_i y_i = 0.$$

The soft margin SVM introduces slack variables and corresponds to the loss function

$$L(f(x), y) = (1 - y_i f(x_i))_+,$$

called hinge loss.

Practical aspects of Hilbert space methods

- Simple mathematical framework
- Clear connection to regularization theory
- Easy to analyze (see later)
- Flexibility by adapting kernel and loss function
- Computationally relatively efficient
- Good performance on real world problems

General Theory of Kernels

[Hein 2003], [Hein 2004], [Hein 2004b]

(Conditionally) positive definite kernels

Definition. A symmetric function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a **positive definite (PD) kernel** if for all $n \geq 1$, all x_1, x_2, \dots, x_n and all c_1, c_2, \dots, c_n

$$\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0.$$

The set of all real valued positive definite kernels on \mathcal{X} is denoted $\mathbb{R}_+^{\mathcal{X} \times \mathcal{X}}$.

Definition. A symmetric function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a **conditionally positive definite (CPD) kernel** if for all $n \geq 1$, all x_1, x_2, \dots, x_n

$$\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$$

for all c_1, c_2, \dots, c_n satisfying $\sum_{i=1}^n c_i = 0$.

Closure properties

Theorem. Given PD/CPD kernels $k_1, k_2 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ the following are also PD/CPD kernels:

$$\begin{aligned}(k_1 + k_2)(x, y) &= k_1(x, y) + k_2(x, y) \\ (\lambda k)(x, y) &= \lambda k_1(x, y) \quad \lambda > 0 \\ k_{1,2}(x, y) &= k_1(x, y) k_2(x, y)\end{aligned}$$

Furthermore, given a sequence of PD/CPD kernels $k_i(x, y)$ converging uniformly to $k(x, y)$, $k(x, y)$ is also PD/CPD.

Theorem. Given PD/CPD kernels $k_1 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $k_2 : \mathcal{X}' \times \mathcal{X}' \rightarrow \mathbb{R}$, the following are PD/CPD kernels on $\mathcal{X} \times \mathcal{X}'$:

$$\begin{aligned}(k_1 \otimes k_2)((x, y)(x', y')) &= k_1(x, y) k_2(x', y') \\ (k_1 \oplus k_2)((x, y)(x', y')) &= k_1(x, y) + k_2(x', y').\end{aligned}$$

Reproducing Kernel Hilbert Spaces

Definition. A **Reproducing Kernel Hilbert Space (RKHS)** on \mathcal{X} is a Hilbert space of functions from \mathcal{X} to \mathbb{R} where all evaluation functionals $\delta_x: \mathcal{H} \rightarrow \mathbb{R}$, $\delta_x(f) = f(x)$ are continuous w.r.t. the topology induced by the norm of \mathcal{H} . Equivalently, for all $x \in \mathcal{X}$ there exists an $M_x < \infty$ such that

$$\forall f \in \mathcal{H}, \quad |f(x)| \leq M_x \|f\|_{\mathcal{H}}.$$

The kernel \leftrightarrow RKHS connection

Theorem. A Hilbert space \mathcal{H} of functions $f: \mathcal{X} \rightarrow \mathbb{R}$ is an RKHS if and only if there exists a function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that

$$\begin{aligned} \forall x \in \mathcal{X} \quad k_x := k(x, \cdot) \in \mathcal{H} \quad \text{and} \\ \forall x \in \mathcal{X} \quad \forall f \in \mathcal{H} \quad \langle f, k_x \rangle = f(x). \end{aligned}$$

If such a k exists, then it is unique and it is a positive definite kernel.

Theorem. If $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a positive definite kernel, then there exists a unique RKHS on \mathcal{X} whose kernel is k .

1. Consider the space of functions spanned by all finite linear combinations

$$\left\{ \sum_{i=1}^n \alpha_i k_{x_i} \mid n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

2. Induce an inner product from $\langle k_x, k_{x'} \rangle = k(x, x')$, which in turn induces a norm $\|\cdot\|$.
3. Complete the space w.r.t. $\|\cdot\|$ to get \mathcal{H} .

Kernel operators

Definition. The operator $K : L_2(\mathcal{X}, \mu) \rightarrow L_2(\mathcal{X}, \mu)$ associated with the kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is defined by

$$(Kf)(x) = \int_{\mathcal{X}} k(x, y) f(y) d\mu(y).$$

Theorem. The operator K is positive, self-adjoint, Hilbert-Schmidt ($\sum \lambda_i^2 < \infty$) and trace-class.

Theorem. (Riesz) If $k \in L_2(\mathcal{X} \times \mathcal{X}, \mu \otimes \mu)$ then there exists an orthonormal system (ϕ_i) in $L_2(\mu)$ such that

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$$

where $\lambda_i \geq 0$ and the sum converges in $L_2(\mathcal{X} \times \mathcal{X}, \mu \otimes \mu)$. Here (ϕ_i) are the eigenvectors of K i.e., $K\phi_i = \lambda_i \phi_i$.

Feature maps

The feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ (satisfying $k(x, x') = \langle \phi(x), \phi(x') \rangle$) is not unique. Important special cases are:

1. Aronszajn map. $\mathcal{H} = \text{RKHS}(k)$ and $\phi: x \mapsto k_x = k(x, \cdot)$.
2. Kolmogorov map. $\mathcal{H} = L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ where μ is a Gaussian measure, $\phi: x \mapsto X_x$ and $k(x, x') = \mathbb{E}[X_x X_{x'}]$. (Gaussian processes)
3. Integral map. For a set T and a measure μ on T , let $\mathcal{H} = L_2(T, \mu)$ and $\phi: x \mapsto (\Gamma_x(t))_T$ and $k(x, x') = \int \Gamma(x, t) \Gamma(x', t) d\mu(t)$. (Bhattacharyya)
4. Riesz map. If $\mathcal{H} = \ell_2(\mathcal{N})$ and $\phi: x \mapsto \sqrt{\lambda_n} \phi_n(x)$ then $k(x, x') = \sum_{i=1}^{\infty} [\phi(x)]_i [\phi(x')]_i$. (Feature map)

Hilbertian Metrics \leftrightarrow CPD kernels

Metric view of SVMs:

$$\begin{array}{lcl} \mathcal{X} & \xrightarrow{k} & \mathcal{H} \longrightarrow \text{max. margin separation} \\ (\mathcal{X}, d) & \xrightarrow{\text{isometric}} & \mathcal{H} \longrightarrow \text{max. margin separation} \end{array}$$

Easy to get d from k :

$$d^2(x, y) = k(x - y, x - y) = k(x, x) + k(y, y) - 2k(x, y).$$

Moreover, $-d^2$ is CPD. For the converse, we can show that all PD kernels are generated by a semi-metric, in the sense that if $-d^2$ is CPD then there exists a function $g: \mathcal{X} \rightarrow \mathbb{R}$ such that

$$k(x, y) = -\frac{1}{2}d^2(x, y) + g(x) + g(y)$$

is PD. Note that this mapping is not one to one: more than one PD kernel corresponds to each CPD metric.

Definition. A semi-metric d on a space \mathcal{X} is Hilbertian if there is an isometric embedding of (\mathcal{X}, d) into some Hilbert space \mathcal{H} .

Theorem. [Schoenberg] a semi-metric d is Hilbertian if and only if $-d^2(x, y)$ is CPD.

Theorem. $k(x, y) = e^{tg(x, y)}$ is PD for all $t > 0$ if and only if g is CPD. [Berg, Christensen, Ressel]

Metric-based SVMs

The SVM optimization problem can be written as

$$\min_{\alpha} -\frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j d^2(x_i, x_j) \quad \text{s. t.} \quad \sum_i y_i \alpha_i = 0, \quad \sum_i \alpha_i = 2, \quad \alpha_i > 0$$

and the solution is

$$f(x) = -\frac{1}{2} \sum_i y_i \alpha_i d^2(x_i, x) + c.$$

The SVM only cares about the metric, not the kernel! [Scholkopf 2000] What about non-Hilbertian metrics? Need separate primal/dual Banach spaces:

$$\begin{aligned} \Phi: (\mathcal{X}, d) &\rightarrow_{\text{isom}} (\overline{D}, \|\cdot\|_{\infty}) & \Psi: \mathcal{X} &\rightarrow E \\ \Phi: x &\mapsto \Phi_x = d(x, \cdot) - d(x_0, \cdot) & \Psi: x &\mapsto \Psi_x = d(\cdot, x) - d(x_0, x) \end{aligned}$$

Giving E the norm

$$\|e\|_E = \inf_{I, (\beta_i)} \left[\sum_{i \in I} |\beta_i| \quad \text{s.t.} \quad e = \sum_{i \in I} \beta_i \Psi_{x_i}, \quad x_i \in \mathcal{X}, \quad |I| < \infty \right]$$

$(\overline{E}, \|\cdot\|_{\overline{E}})$ is the topological dual of $(\overline{D}, \|\cdot\|_{\overline{D}})$. The analog of the SVM is

$$\inf_{m \in \mathbb{N}, (x_i)_{i=1}^m, b} \sum_{i=1}^m |\beta_i| = 1 \quad \text{s.t.} \quad y_j \left[b + \sum_{i=1}^m \beta_i (d(x_j, x_i) - d(x_i, x_0)) \right] \geq 1.$$

(Max. distance between convex hulls \leftrightarrow max. margin hyperplane.) No representer theorem!

Fuglede's Theorem

Definition. A symmetric function k is γ -homogeneous if $k(cx, cy) = c^\gamma k(x, y)$.

Theorem. A symmetric function $d: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $d(x, x) \Leftrightarrow x = 0$ is a 2γ -homogeneous continuous Hilbertian metric on \mathbb{R}_+ if and only if there exists a (necessarily unique) non-zero bounded measure $\rho \geq 0$ on \mathbb{R}_+ such that

$$d^2(x, y) = \int_{\mathbb{R}_+} |x^{\gamma+i\lambda} - y^{\gamma-i\lambda}|^2 d\rho(\lambda).$$

Corollary. A symmetric function $k: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $k(x, x) \Leftrightarrow x = 0$ is a 2γ -homogeneous continuous PD kernel on \mathbb{R}_+ if and only if there exists a (necessarily unique) non-zero bounded measure $\kappa \geq 0$ on \mathbb{R}_+ such that

$$d^2(x, y) = \int_{\mathbb{R}_+} x^{\gamma+i\lambda} y^{\gamma-i\lambda} d\kappa(\lambda).$$

General Covariant Kernels on $\mathcal{M}_+^1(\mathcal{X})$

Theorem. Let P and Q be two probability measures on \mathcal{X} , μ a dominating measure of P and Q , and $d_{\mathbb{R}_+}$ a $1/2$ -homogeneous Hilbertian metric on \mathbb{R}_+ . Then $D_{\mathcal{M}_+^1(\mathcal{X})}$

$$D_{\mathcal{M}_+^1(\mathcal{X})}^2(P, Q) = \int_{\mathcal{X}} d_{\mathbb{R}_+}^2(p(x), q(x)) d\mu(x)$$

is a Hilbertian metric on $\mathcal{M}_+^1(\mathcal{X})$ that is independent of μ .

The corresponding kernels are:

$$K_{\frac{1}{2}|1}(P, Q) = \int_{\mathcal{X}} \sqrt{p(x)q(x)} d\mu(x) \quad (\text{Bhattacharyya})$$

$$K_{1|-1}(P, Q) = \int_{\mathcal{X}} \frac{p(x)q(x)}{p(x) + q(x)} d\mu(x)$$

$$K_{1|1}(P, Q) = -\frac{1}{\log 2} \int_{\mathcal{X}} \left[p(x) \log \left(\frac{p(x)}{p(x) + q(x)} \right) + q(x) \log \left(\frac{q(x)}{p(x) + q(x)} \right) \right] d\mu(x)$$

$$K_{\infty|1}(P, Q) = \int_{\mathcal{X}} \min [p(x), q(x)] d\mu(x)$$

Sequences

[Hausser 1999] [Watkins 1999] [Leslie 2003] [Cortes 2004]

Convolution kernels

Assume that each $x \in \mathcal{X}$ can be decomposed into “parts” described by the relation $R(x_1, x_2, \dots, x_D, x)$ with $\vec{x} = x_1, x_2, \dots, x_D \in \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_D$ in possibly multiple ways $R^{-1}(x) = \{\vec{x}_1, \vec{x}_2, \dots\}$. Given kernels $k_i: \mathcal{X}_i \times \mathcal{X}_i \rightarrow \mathbb{R}$, their convolution kernel is defined

$$k(x, y) = (k_1 \star k_2 \star \dots \star k_D)(x, y) = \sum_{\vec{x} \in R^{-1}(x), \vec{y} \in R^{-1}(y)} \prod_{d=1}^D k_d(x_d, y_d).$$

E.g. Gaussian RBF kernel btw. $x = (x_1, x_2, \dots, x_D)$ and $y = (y_1, y_2, \dots, y_D)$

$$k(x, y) = \prod_{d=1}^D k_d(x, y) \quad k_d(x, y) = \exp(-(x_d - y_d)^2 / (2\sigma^2)).$$

E.g. The ANOVA kernel for $\mathcal{X} = S^D$ is

$$k(x, y) = \sum_{1 \leq i_1 \leq \dots \leq i_d \leq n} \prod_{d=1}^D k_{i_d}(x_{i_d}, y_{i_d}).$$

Iterated convolution kernels

A P -kernel is a kernel that is also a probability distribution on $\mathcal{X} \times \mathcal{X}$, i.e., $k(x, y) \geq 0$ and $\sum_{x, y} k(x, y) = 1$.

The relationship R between x and its parts is a function if for every \vec{x} there is an $x \in \mathcal{X}$ such that $R^{-1}(x) = \vec{x}$. Assume that R is a finite function that is also associative in the sense that if $x_1 \circ x_2 = x$ denotes $R(x_1, x_2, x)$ then $(x_1 \circ x_2) \circ x_3 = x_1 \circ (x_2 \circ x_3)$. Defining $k^{(r)} = k \star k^{(r-1)}$, the γ -infinite iteration of k is

$$k_{\gamma}^{\star} = (1 - \gamma) \sum_{r=1}^{\infty} \gamma^{r-1} k^{(r)}.$$

Substitution kernel: $k_1(x, y) = \sum_{a \in \mathcal{A}} p(a) k_a(x, y)$

Insertion kernel: $k_2(x, y) = g(x)g(y)$

REGular string kernel:

$$k(x, y) = \gamma k_2 \star (k_1 \star k_2)_{\gamma}^{\star} + (1 - \gamma) k_2.$$

Watkins' Substring Kernels

We say that u is a substring of s indexed by $\mathbf{i} = i_1, i_2, \dots, i_{|u|}$ if $u_j = s_{i_j}$. We denote this relationship by $u = s[\mathbf{i}]$ and let $l(\mathbf{i}) = i_{|u|} - i_1 + 1$. For some $\lambda > 0$ the kernel corresponding to the explicit feature mapping $\phi_u(s) = \sum_{\mathbf{i}:s[\mathbf{i}]=u} \lambda$ is

$$k_n(s, t) = \sum_{u \in \Sigma^n} \sum_{\mathbf{i}:u=s[\mathbf{i}]} \sum_{\mathbf{j}:u=t[\mathbf{j}]} \lambda^{l(\mathbf{i})+l(\mathbf{j})}.$$

Defining

$$k'_p(s, t) = \sum_{u \in \Sigma^p} \sum_{\mathbf{i}:u=s[\mathbf{i}]} \sum_{\mathbf{j}:u=t[\mathbf{j}]} \lambda^{l(\mathbf{i})+l(\mathbf{j})}.$$

a recursive computation is possible by

$$k'_0(s, t) = 1$$

$$k'_p(s, t) = 0 \quad \text{if} \quad |s| < p \quad \text{or} \quad |t| < p$$

$$k_p(s, t) = 0 \quad \text{if} \quad |s| < p \quad \text{or} \quad |t| < p$$

$$k'_p(sx, t) = \lambda k'_p(s, t) + \sum_{j:t_j=x} k'_{i-1}(s, t[1:j-1]) \lambda^{|t|-j+2}$$

$$k_n(sx, t) = k_n(s, t) + \sum_{j:t_j=x} k'_{n-1}(s, t[1:j-1]) \lambda^2$$

Mismatch Kernels

Mismatch feature map:

$$[\phi_{(k,m)}^{\text{Mismatch}}(x)]_{\beta} = \sum_{\alpha \in \Sigma^k, \alpha \sqsubset x} I(\beta \in N_{(k,m)}(\alpha)) \quad \beta \in \Sigma^k$$

Restricted gappy feature map:

$$[\phi_{(g,k)}^{\text{Gappy}}(x)]_{\beta} = \sum_{\alpha \in \Sigma^k, \alpha \sqsubset x} I(\alpha \in G_{(g,k)}(\beta)) \quad \beta \in \Sigma^k$$

Substitution feature map: as mismatch feature map, but

$$N_{(k,\sigma)}(\alpha) = \left\{ \beta = b_1 b_2 \dots b_k \in \Sigma^k : - \sum_{i=1}^k \log P(a_i | b_i) < \sigma \right\}$$

Computing these kernels using a prefix trie gives $O(g^{g-k+1} (|x| + |y|))$ algorithms.

Finite State Transducers

Alphabets: Σ, Δ

Semiring: K (operations \oplus, \otimes)

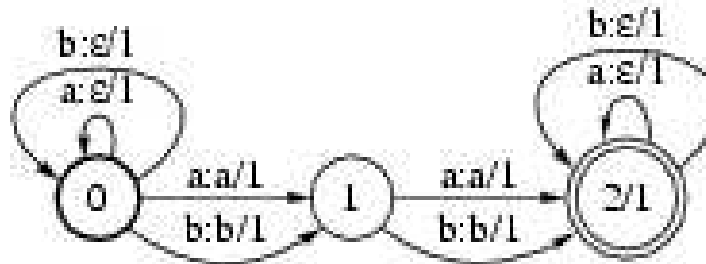
Edges: e_i

Weights: $w(e)$

Final weights: λ_i

Transducer: $\Sigma^* \times \Delta^* \rightarrow K$

Set of Paths: $P(x, y) \quad x \in \Sigma^*, y \in \Delta^*$



Total weight assigned to pair of input/output strings x and y (regulated transducer):

$$[[T]](x, y) = \bigoplus_{\pi \in P(x, y)} \lambda(\pi) \otimes \bigotimes_{e \in \pi} w(e)$$

Operations on transducers:

$$[[T_1 \oplus T_2]](x, y) = [[T_1]](x, y) \oplus [[T_2]](x, y) \quad \text{(parallel)}$$

$$[[T_1 \otimes T_2]](x, y) = \bigoplus_{x=x_1x_2 \quad y=y_1y_2} [[T_1]](x_1, y_1) \otimes [[T_2]](x_2, y_2) \quad \text{(series)}$$

$$[[T^*]](x, y) = \bigoplus_{n=0}^{\infty} [[T^n]](x, y) \quad \text{(closure)}$$

$$[[T_1 \circ T_2]](x, y) = \bigoplus_{z \in \Delta^*} [[T_1]](x, z) \otimes [[T_2]](z, y) \quad \text{(composition)}$$

Rational Kernels

Definition. A positive definite function $k : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$ is called a **rational kernel** if there exists a transducer T and a function $\psi : K \rightarrow \mathbb{R}$ such that

$$k(x, y) = \psi (\llbracket T \rrbracket (x, y)) .$$

Naturally extends to a kernel over weighted automata.

Theorem. Rational kernels are closed under \oplus sum, \otimes product, and $*$ Kleene closure.

Theorem. Assume that $T \circ T^{-1}$ is regulated and ψ is a semiring morphism. Then $k(x, y) = \psi (\llbracket T \circ T^{-1} \rrbracket (x, y))$ is a rational kernel.

Theorem. There exist $O(|T| |x| |y|)$ algorithms for computing $k(x, y)$.

Spectral Kernels

[Kondor 2002], [Belkin 2002], [Smola 2003]

The Laplacian

Discrete case (graphs).

$$\Delta_{ij} = \begin{cases} w_{ij} & i \sim j \\ -\sum_k w_{ik} & i = j \\ 0 & \text{otherwise} \end{cases} \quad \text{or} \quad \tilde{\Delta} = D^{-1/2} \Delta D^{-1/2}.$$

Continuous case (Riemannian manifolds)

$$\Delta: L_2(\mathcal{M}) \rightarrow L_2(\mathcal{M})$$

$$\Delta = \frac{1}{\sqrt{\det g}} \sum_{ij} \delta_i \sqrt{\det g} g^{ij} \delta_j$$

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \dots + \frac{\partial^2}{\partial x_D^2} \quad \text{on} \quad \mathbb{R}^D$$

The heat kernel (diffusion kernel)

$$K = e^{t\Delta} = \lim_{n \rightarrow \infty} \left(I + \frac{t\Delta^n}{n} \right) \quad k(x, x') = \langle \delta_x, K\delta_{x'} \rangle$$

Δ self-adjoint $\Rightarrow k$ positive definite. Well studied and natural interpretations on many different objects. On \mathbb{R}^D we get back the familiar Gaussian RBF

$$k(x, x') = \frac{1}{(2\pi\sigma^2)^{D/2}} e^{-|x-x'|^2/(2\sigma^2)}.$$

On p -regular trees as a function of distance d

$$k(i, j) = \frac{2}{\pi(p-1)} \int_0^\pi \frac{e^{-\beta\left(1 - \frac{2\sqrt{p-1}}{p} \cos x\right)} \sin x [(p-1) \sin(d+1)x - \sin(d-1)x]}{p^2 - 4(p-1) \cos^2 x} dx.$$

Approximating the heat kernel on a data manifold

The assumption is that our data lives on a manifold \mathcal{M} embedded in \mathbb{R}^n . Given $X = x_1, x_2, \dots, x_m$ (labeled and unlabeled data points) sampled from \mathcal{M} , the graph Laplacian approximates the Laplace operator on \mathcal{M} in the sense that

$$\langle f, \Delta g \rangle_{L_2(\mathcal{M})} \approx \langle f|_X, \Delta_{\text{graph}} g|_X \rangle.$$

The graph Laplacian $W - D$ can be constructed in different ways:

1. $w_{ij} = 1$ if $\|x_i - x_j\| < \epsilon$, otherwise $w_{ij} = 0$
2. $w_{ij} = 1$ if i is amongst the k nearest neighbors of j or j is amongst the k nearest neighbors of i , otherwise $w_{ij} = 0$
3. $w_{ij} = \exp(-\|x_i - x_j\|^2 / (2\sigma^2))$

First few eigenvectors of Δ provide natural basis for low-dimensional projection of \mathcal{M} .

Other spectral kernels

The exponential map is not the only way to get a regularization operator (kernel) from the Laplacian. General form:

$$\langle f, f \rangle = \langle f, P^* P f \rangle^{L_2} = \sum_i r(\lambda_i) \langle f, \phi_i \rangle_{L_2} \langle \phi_i, f \rangle_{L_2}$$

where ϕ_1, ϕ_2, \dots is an eigensystem of Δ with corresponding eigenvalues $\lambda_1, \lambda_2, \dots$

$r(\lambda) = 1 + \sigma^s \lambda$	regularized Laplacian
$r(\lambda) = \exp(\sigma^2/(2\lambda))$	diffusion kernel
$r(\lambda) = (aI - \lambda)^{-p}$	p -step random walk
$r(\lambda) = (\cos \lambda\pi/4)$	inverse cosine

The Laplacian is the essentially unique linear operator on $L_2(\mathcal{X})$ invariant under the group of isometries of a general metric space \mathcal{X} . All kernels invariant in the same sense can be derived from Δ by a suitable choice of function r .

Kernels on Distributions

[Lafferty 2002] [Jebara 2003] [Kondor 2003]

Information Diffusion Kernels

A d -dimensional parametric family $\{p_\theta(\cdot), \theta \in \Theta \subset \mathbb{R}^d\}$ gives rise to a Riemannian manifold with Fisher metric

$$g_{ij}(\theta) = \mathbb{E}[(\partial_i \ell_\theta)(\partial_j \ell_\theta)] = \int_{\mathcal{X}} (\partial_i \log p(x|\theta)) (\partial_j \log p(x|\theta)) p(x|\theta) dx =$$

$$4 \int_{\mathcal{X}} \left(\partial_i \sqrt{p(x|\theta)} \right) \left(\partial_j \sqrt{p(x|\theta)} \right) dx$$

In terms of the metric, the Laplacian is

$$\Delta = \frac{1}{\sqrt{\det g}} \sum_{ij} \delta_i \sqrt{\det g} g^{ij} \delta_j$$

which we can exponentiate to get the diffusion kernel. The general form is

$$k_t(x, y) = (4\pi t)^{-d/2} \exp\left(-\frac{d^2(x, y)}{4t}\right) \sum_{i=1}^N \psi_i(x, y) t^i + O(t^N)$$

The information geometry of the multinomial is isometric to the positive quadrant of the hypersphere where

$$k_t(\theta, \theta') = (4\pi t)^{-d/2} \exp\left(-\frac{1}{t} \arccos^2\left(\sum_{i=1}^{d+1} \sqrt{\theta_i \theta'_i}\right)\right).$$

Probability Product Kernels

For p and p' distributions on \mathcal{X} and $\rho > 0$

$$k(p, p') = \int_{\mathcal{X}} p(x)^\rho p'(x)^\rho dx = \langle p^\rho, p'^\rho \rangle_{L_2}$$

Bhattacharyya ($\rho = 1/2$):

$$k(p, p') = \int \sqrt{p(x)} \sqrt{p'(x)} dx$$

Satisfies $k(p, p) = 1$ and related to Hellinger's distance

$$H(p, p') = \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{p'(x)} \right)^2 dx$$

by $H(p, p') = \sqrt{2 - 2k(p, p')}$.

Expected likelihood kernel ($\rho = 1$):

$$K(x, x') = \int p(x) p'(x) dx = \mathbf{E}_p[p'(x)] = \mathbf{E}_{p'}[p(x)].$$

Probability Product Kernels for Exponential Families

Gaussians:

$$k_\rho(p, p') = \int_{\mathbb{R}^D} p(x)^\rho p'(x)^\rho dx =$$

$$(2\pi)^{(1-2\rho)D/2} \rho^{-D/2} |\Sigma^\dagger|^{1/2} |\Sigma|^{-\rho/2} |\Sigma'|^{-\rho/2} \exp\left(-\frac{\rho}{2} \left(\mu^T \Sigma^{-1} \mu + \mu'^T \Sigma'^{-1} \mu' - \mu^{\dagger T} \Sigma^\dagger \mu^\dagger\right)\right)$$

where $\Sigma^\dagger = (\Sigma^{-1} + \Sigma'^{-1})^{-1}$ and $\mu^\dagger = \Sigma^{-1} \mu + \Sigma'^{-1} \mu'$

Bernoulli:

$$p(x) = \prod_{d=1}^D \gamma_d^{x_d} (1 - \gamma_d)^{1-x_d} \quad K_\rho(p, p') = \prod_{d=1}^D [(\gamma_d \gamma'_d)^\rho + (1 - \gamma_d)^\rho (1 - \gamma'_d)^\rho]$$

Multinomial ($\rho = 1/2$):

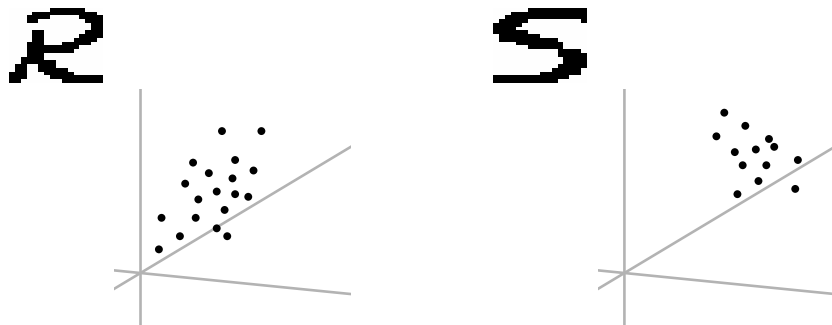
$$K(p, p') = \sum \frac{s!}{x_1! x_2! \dots x_D!} \prod_{d=1}^D (\alpha_d \alpha'_d)^{x_d/2} = \left[\sum_{d=1}^D (\alpha_d \alpha'_d)^{1/2} \right]^s$$

Gamma:

$$p(x) = \frac{1}{\Gamma(\alpha) \beta^\alpha} x^{\alpha-1} e^{-x/\beta} \quad k_\rho(p, p') = \frac{\Gamma(\alpha^\dagger) \beta^{\alpha^\dagger}}{\left[\Gamma(\alpha) \beta^\alpha \Gamma(\alpha') \beta'^{\alpha'} \right]^\rho}$$

Feature Space Bhattacharyya Kernels

Base kernel (e.g. Gaussian RBF) maps points to feature space



Kernel between examples, $K(x, x')$ is computed as feature space Bhattacharyya between two fitted Gaussians with mean and covariance

$$\hat{\mu} = \frac{1}{k} \sum_{i=1}^k \Phi(x_i) \quad \hat{\Sigma}_{\text{reg}} = \sum_{l=1}^r v_l \lambda_l v_l^\top + \eta \sum_i \zeta_i \zeta_i^\top$$

Tropical Geometry of Graphical Models

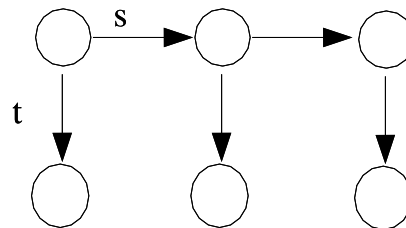
[Pachter 2004a], [Pachter 2004b]

Tropical Geometry and Bayesian Networks

Parameters: s_1, s_2, \dots, s_d

Observations: $\sigma_1, \sigma_2, \dots, \sigma_m$

Mapping: $f: \mathbb{R}^d \rightarrow \mathbb{R}^m$



$$f_{\sigma_1, \sigma_2, \dots, \sigma_m}(s) = p(\sigma_1, \sigma_2, \dots, \sigma_m | s) = \sum_{h_1, \dots, h_k} p(\sigma_1, \sigma_2, \dots, \sigma_m | h_1, h_2, \dots, h_k, s)$$

e.g., for 3-state HMM

$$f_{\sigma_1, \sigma_2, \sigma_3} = s_{00}s_{00}t_{0\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{00}s_{01}t_{0\sigma_1}t_{0\sigma_2}t_{1\sigma_3} + s_{01}s_{10}t_{0\sigma_1}t_{1\sigma_2}t_{0\sigma_3} + s_{01}s_{11}t_{0\sigma_1}t_{1\sigma_2}t_{1\sigma_3} + \\ s_{10}s_{00}t_{1\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{10}s_{01}t_{1\sigma_1}t_{0\sigma_2}t_{1\sigma_3} + s_{11}s_{10}t_{1\sigma_1}t_{1\sigma_2}t_{0\sigma_3} + s_{11}s_{11}t_{1\sigma_1}t_{1\sigma_2}t_{1\sigma_3}$$

Tropicalization

Tropicalization to find max. log likelihood sequence:

$(+, \times)$ -semiring \rightarrow $(\min, +)$ -semiring

$f \rightarrow \delta = -\log f$

$s_{ij} \rightarrow u_{ij} = -\log u_{ij}$

$t_{ij} \rightarrow v_{ij} = -\log v_{ij}$

e.g., for 3-state HMM we get Viterbi path by

$$\delta_{\sigma_1, \sigma_2, \dots, \sigma_m} = \min_{h_1, h_2, h_3} [u_{h_1 h_2} + u_{h_2 h_3} + v_{h_1 \sigma_1} + v_{h_2 \sigma_2} + v_{h_3 \sigma_3}]$$

Let (\vec{a}_i) be vectors of exponents of the parameters corresponding to different settings of the hidden variables. Then $\delta_{\sigma_1, \sigma_2, \dots, \sigma_m} = \min_i [\vec{a}_i \cdot \vec{u}]$. The ML solution changes when $(\vec{a}_i - \vec{a}_j) \cdot \vec{u} = 0$ for $i \neq j$. Feasible values of \vec{a}_i are vertices of the **Newton polytope** of f and δ is linear in each **normal cone** of the Newton polytope.

The Algebraic Geometry

Polytope propagation:

$$\begin{aligned} \text{NP}(f \cdot g) &= \text{NP}(f_1) + \text{NP}(f_2) & A + B &= \{a + b \mid a \in A, b \in B\} \\ \text{NP}(f + g) &= \text{NP}(f_1) \cup \text{NP}(f_2). \end{aligned}$$

Can run the sum-product algorithm with polytopes!

Each vertex of $\text{NP}(f_\sigma)$ corresponds to a ML solution. Each vertex of $\text{NP}(f_\sigma)$ corresponds to an inference function $\sigma \rightarrow h$. Key observation:

$$\#\text{vertices}(\text{NP}(f_\sigma)) \leq \text{const.} \cdot E^{d(d-1)/(d+1)}.$$

Generalization bounds

[Mendelson 2003] [Bousquet 2004] [Lugosi 2003] [Bartlett 2003] [Bousquet 2003]

The Classical Approach: Union Bound

Recall, we are interested in bounding

$$\sup_{f \in \mathcal{F}} (Pf - P_n f)$$

where $\mathcal{F} = \{ L \circ f \mid f \in \mathcal{F}_{\text{orig}} \}$.

For a fixed f , assuming $f(x) \in [a, b]$, by Hoeffding

$$\mathbb{P}[|Pf - P_n f| > \epsilon] \leq 2 \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right), \quad \mathbb{P}\left[|Pf - P_n f| > (b-a) \sqrt{\frac{\log \frac{2}{\delta}}{2n}}\right] \leq \delta.$$

Now taking union over all $f \in \mathcal{F}$ when \mathcal{F} is finite,

$$\sup_{f \in \mathcal{F}} |Pf - P_n f| \leq \sqrt{\frac{\log |\mathcal{F}| + \log \frac{1}{\delta}}{2n}}$$

with probability $1 - \delta$.

“Ockham’s Razor” bound

Reweighting by $p(f)$ s.t. $\sum_{f \in \mathcal{F}} p(f) = 1$ we can extend above to the countably infinite case. With probability $1 - \delta$

$$|Pf - P_n f| \leq \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\delta}}{2n}}$$

simultaneously for all $f \in \mathcal{F}$.

A related idea is the PAC-Bayes bound for binary stochastic classifiers described by a distribution $Q(x)$:

$$\sup_Q \text{KL}(P_n[Q] \| P[Q]) \leq \frac{1}{m} [\text{KL}(Q \| P) + \log \frac{m+1}{\delta}]$$

with probability $1 - \delta$ for any prior P . A particular application is the margin-dependent PAC Bayes bound for stochastic hyperplane classifiers.

Alternative Measures of Generalization Error

1. Mendelson:

$$\mathbb{P}[\exists f \in \mathcal{F} : Pf < \epsilon, P_n f \geq 2\epsilon]$$

2. Normalization (Vapnik):

$$\mathbb{P}\left[\frac{Pf - P_n f}{\sqrt{Pf}} < \epsilon\right]$$

3. Localized Rademacher complexities

4. Algorithmic stability

5. ...

Vapnik-Chervonenkis Theory

A set x_1, x_2, \dots, x_m is shattered by \mathcal{F} if for every $I \subset \{1, 2, \dots, m\}$, there is a function $f_I \in \mathcal{F}$ such that $f(x_i) = \mathbf{1}(i \in I)$. The VC-dimension is defined

$$d = VC(\mathcal{F}) = \max_{X \subset \mathcal{X}} |X| \quad \text{such that } X \text{ is shattered by } \mathcal{F}.$$

Defining the coordinate projection of \mathcal{F} on X as $P_X \mathcal{F} = \{ (f(x_i))_{x_i \in X} \mid f \in \mathcal{F} \}$, the growth function is $\Pi(n) = \max_{X \subset \mathcal{X}} |P_X \mathcal{F}|$. By the Sauer-Shelah Lemma $\Pi(n) \leq \left(\frac{en}{d}\right)^d$.

A set x_1, x_2, \dots, x_m is ϵ -shattered by \mathcal{F} if there is some function $s: \mathcal{X} \rightarrow \mathbb{R}$ such that for every $I \subset \{1, 2, \dots, m\}$, there is a function $f_I \in \mathcal{F}$ such that

$$f(x_i) \geq s(x_i) + \epsilon \quad \text{if } i \in I \quad f(x_i) \leq s(x_i) - \epsilon \quad \text{if } i \notin I.$$

The fat-shattering dimension is $d_\epsilon(\mathcal{F}) = \max_{X \subset \mathcal{X}} |X|$ such that X is ϵ -shattered by \mathcal{F} .

A classical result from VC-theory is that for binary valued classes

$$\sup_{f \in \mathcal{F}} [Pf - P_n f] \leq 2 \sqrt{\frac{\log \Pi(2n) + \log \frac{2}{\delta}}{(n/2)}}$$

Symmetrization and Rademacher Averages

The Rademacher average of \mathcal{F} is

$$R_n(\mathcal{F}) = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right]$$

where the σ_i are $\{-1, +1\}$ -valued random variables with $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

By Vapnik's classical symmetrization argument

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} [Pf - P_n f] \right] \leq 2R_n(\mathcal{F})$$

Strategy: investigate the concentration of $R_n(\mathcal{F})$ about its mean, as well as the concentration of $\sup_{f \in \mathcal{F}} [Pf - P_n f]$ about its mean. Example of a resulting bound (from McDiarmid):

$$\sup_{f \in \mathcal{F}} [Pf - P_n f] \leq 2R_n(\mathcal{F}) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}$$

with probability $1 - \delta$. For kernel classes

$$R_n(\mathcal{F}) \leq \frac{\gamma}{n} \left(\sum_{i=1}^n k(x_i, x_i) \right)^{1/2}$$

where $\gamma = \|\hat{f}\|$ and \hat{f} is the function returned by our algorithm.

Classical Concentration Inequalities

Markov: for any r.v. $X \geq 0$

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}X}{t}.$$

Chebyshev:

$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq \frac{\text{Var}(X)}{\text{Var}(X) + t^2}.$$

Hoeffding: ($|X_i| < c$)

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X\right| > \epsilon\right] \leq 2 \exp\left(-\frac{n\epsilon^2}{2c^2}\right)$$

Bernstein: ($|X_i| < c$)

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X\right| > \epsilon\right] < \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right)$$

Tools: Chernoff's bounding method, entropy method

Uniform Concentration Inequalities

Talagrand's inequality. Let $Z = \sup_{f \in F} [Pf - P_n f]$, $b = \sup_{x \in \mathcal{X}} Z$ and $v = \sup_{f \in F} P(f^2)$. Then there is an absolute constant C such that with probability $1 - \delta$

$$Z \leq 2\mathbb{E}Z + C \left(v \sqrt{\log \frac{1}{\delta}} + b \log \frac{1}{\delta} \right).$$

Bousquet's inequality. Under the same conditions as above, with probability $1 - \delta$

$$Z \geq \inf_{\alpha > 0} \left[(1 + \alpha)\mathbb{E}[Z] + \sqrt{\frac{2v}{n}} + \left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{b \log \frac{1}{\delta}}{n} \right].$$

Surrogate Loss functions

In classification, ultimate measure of loss is 0-1 loss. Instead algorithms often minimize a surrogate loss $L(f(x), y) = \phi(yf(x))$.

	$\phi(\alpha)$	
exponential	$e^{-\alpha}$	$1 - \sqrt{1 - \theta^2}$
logistic	$\ln(1 + e^{-2\alpha})$	θ
quadratic	$(1 - \alpha)^2$	θ^2

$$\text{Risk: } R[f] = \mathbb{E} [\mathbf{1}_{\text{sgn}(f(x)) \neq y}] \quad R^* = \inf_f R[f]$$

$$\phi\text{-risk: } R_\phi[f] = \mathbb{E} [\phi(yf(x))] \quad R_\phi^* = \inf_f R_\phi[f]$$

What is the relationship between $R[f] - R^*$ and $R_\phi^*[f] - R_\phi^*$?

Classification calibration

$$\eta(x) = \mathbb{P}[y = 1 \mid x]$$

Optimal conditional ϕ -risk:

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} (\eta \phi(\alpha) - (1 - \eta) \phi(-\alpha))$$

Optimal incorrect conditional ϕ -risk:

$$H^-(\eta) = \inf_{\alpha(2\eta-1) \leq 0} (\eta \phi(\alpha) - (1 - \eta) \phi(-\alpha))$$

Definition: ϕ is classification-calibrated if

$$H^-(\eta) > H(\eta).$$

ψ -transform

$\psi: [0, 1] \rightarrow \mathbb{R}^+$ defined $\psi = \tilde{\psi}^{**}$ where

$$\tilde{\psi}(\theta) = H^-((1+\theta)/2) - H((1+\theta)/2).$$

Theorem: For any nonnegative ϕ and measurable f

$$\psi(R[f] - R^*) \leq R_\phi[f] - R_\phi^*.$$

and for any $\theta \in [0, 1]$ there is a function $f: \mathcal{X} \rightarrow \mathbb{R}$ such that this inequality is ϵ -tight.

Theorem: If ϕ is convex, then it is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.

Theorem: If ϕ is convex, then it is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.

References

Hilbert Space Methods

[Scholkopf 2001] B. Scholkopf and A. Smola. **Learning with Kernels**

General Theory of RKHSs

[Hein 2003] M. Hein and O. Bousquet. **Maximal Margin Classification for Metric Spaces**

[Hein 2004] M. Hein and O. Bousquet. **Kernels, Associated Structures and Generalizations.**

[Hein 2004b] M. Hein and O. Bousquet. **Hilbertian Metrics and Positive Definite Kernels on Probability Measures.**

Regularization Theory

[Girosi 1995] Girosi, F., M. Jones, and T. Poggio. **Regularization Theory and Neural Network Architectures.**

[Smola 1998] A. Smola and B. Scholkopf. **From Regularization Operators to Support Vector Kernels Advances**

Tropical Geometry of Graphical Models

[Pachter 2004a] L. Pachter and B. Sturmfels. **Tropical Geometry of Statistical Models**

[Pachter 2004b] L. Pachter and B. Sturmfels. **Parametric Inference for Biological Sequence Analysis**

Sequences

[Hausssler 1999] D. Hausssler. **Convolution Kernels on Discrete Structures**

[Watkins 1999] Chris Watkins. **Dynamic Alignment Kernels.**

[Leslie 2003] Christina Leslie and Rui Kuang. **Fast Kernels for Inexact String Matching**

[Cortes 2004] Corinna Cortes, Patrick Haffner, and Mehryar Mohri. **Rational Kernels: Theory and Algorithms.**

Spectral Kernels

[Kondor 2002] R. I. Kondor and J. Lafferty. **Diffusion Kernels on Graphs and Other Discrete Input Spaces.**

[Belkin 2002] M. Belkin and P. Niyogi. **Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.**

[Smola 2003] A. Smola and R. Kondor. **Kernels and Regularization on Graphs.**

Generalization Bounds

[Mendelson 2003] S. Mendelson. **A few notes on Statistical Learning Theory.**

[Bousquet 2004] O. Bousquet, S. Boucheron, and G. Lugosi. **Introduction to statistical learning theory.**

[Lugosi 2003] G. Lugosi. **Concentration-of-measure inequalities.**

[Bartlett 2003] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. **Large margin classifiers: convex loss, low noise, and convergence rates.**

[Bartlett 2004] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. **Local Rademacher complexities.**

[Bousquet 2003] Olivier Bousquet. **New Approaches to Statistical Learning Theory.**

[Langford 2002] John Langford and John Shawe-Taylor. **PAC-Bayes and Margins.**