Statistical Techniques in Robotics (16-831, F08) Lecture #23 (Nov 11, 2008)

Kernel Methods / Functional Gradient Descent

Lecturer: Drew Bagnell *Scribe: Daniel Munoz*

1 Goal

The high-level idea is to learn non-linear models using the same gradient-based approach used to learn linear models. Hopefully this will result in better models that improve classification.

2 Review

- Ultimately, we wish to learn a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that assigns a meaningful score given a data point. E.g. in binary classification, we would like $f(\cdot)$ to return positive and negative values, given positive and negative samples, respectively.
- A kernel $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ intuitively measures the *correlation* between $f(\mathbf{x_i})$ and $f(\mathbf{x_j})$. Considering a matrix $\mathbf{K}$ with entries $K_{ij} = K(\mathbf{x_i}, \mathbf{x_j})$, then matrix $\mathbf{K}$ must satisfy the properties:
  - $\mathbf{K}$ is symmetric ($K_{ij} = K_{ji}$)
  - $\mathbf{K}$ is positive-definite ($\forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \neq \mathbf{0}, \mathbf{x}^\mathbf{T}\mathbf{K}\mathbf{x} > 0$)

  Hence, a valid kernel is the inner product: $K_{ij} = \langle \mathbf{x_i}, \mathbf{x_j} \rangle$.
- A function can be considers as a weighted composition of many kernels centered at various locations $\mathbf{x_i}$:

$$f(\cdot) = \sum_{i=1}^{Q} \alpha_i K(\mathbf{x_i}, \cdot), \tag{1}$$

  where Q is the number of kernels that compose $f(\cdot)$ and $\alpha_i \in \mathbb{R}$ is each kernel's associated weight.
  - All functions $f(\cdot)$ with kernel K that satisfy the above properties and can be written in the form of Equation 1 are said to lie in a *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}_K$: $f \in \mathcal{H}_K$
  - The inner-product of two functions f and g is defined as

$$\langle f, g \rangle = \sum_{i=1}^{Q} \sum_{j=1}^{P} \alpha_i \beta_j K(\mathbf{x_i}, \mathbf{x_j}) = \alpha^\mathbf{T} \mathbf{K} \beta, \tag{2}$$

    where $\alpha \in \mathbb{R}^Q$ and $\beta \in \mathbb{R}^P$ are the kernel coefficients for f and g, respectively.
    * By definition, the following property holds: $\langle K(\mathbf{x_i}, \cdot), K(\cdot, \mathbf{x_j}) \rangle = K(\mathbf{x_i}, \mathbf{x_j})$

  * The reproducing property is observed by taking the inner-product of a function with a kernel: $\langle f, K(\mathbf{x_j}, \cdot)\rangle = \langle \sum_{i=1}^{Q} \alpha_i K(\mathbf{x_i}, \cdot), K(\cdot, \mathbf{x_j})\rangle = \sum_{i=1}^{Q} \alpha_i \langle K(\mathbf{x_i}, \cdot), K(\cdot, \mathbf{x_j})\rangle = \sum_{i=1}^{Q} \alpha_i K(\mathbf{x_i}, \mathbf{x_j}) = f(\mathbf{x_j})$
  * Note that due to positive-definite constraint, the squared norm of a function f is always positive when $\alpha \neq \mathbf{0}$. ($||f||^2 = \langle f, f\rangle = \alpha^{\mathbf{T}}\mathbf{K}\alpha > 0$)

- A *functional* $F : f \to \mathbb{R}$ is a function of functions $f \in \mathcal{H}_K$. Examples:
  - $F[f] = ||f||^2$
  - $F[f] = (f(x) - y)^2$
  - $F[f] = \frac{\lambda}{2}||f||^2 + \sum_i (f(x_i) - y_i)^2$

- A functional gradient $\nabla F[f]$ is defined implictly as the linear term of the change in a function due to a small perturbation ϵ in its input: $F[f + \epsilon g] = F[f] + \epsilon\langle \nabla F[f], g\rangle + O(\epsilon^2)$
  - Example: $\nabla F[f] = \nabla ||f||^2 = 2f$

$$\begin{aligned} F[f + \epsilon g] &= \langle f + \epsilon g, f + \epsilon g\rangle \\ &= ||f|| + 2\langle f, \epsilon g\rangle + \epsilon^2 ||g|| \\ &= ||f|| + \epsilon\langle 2f, g\rangle + O(\epsilon^2) \end{aligned}$$

3 More functional gradients

- Consider *differentiable* functions $C : \mathbb{R} \to \mathbb{R}$ that are functions of functionals G, $C(G[f])$. We will be minimizing these (cost) functions in the near future.

- The derivative of these functions follows the chain rule: $\nabla C(G[f]) = C'(G[f])\nabla G[f]$
  - Example: If $C = (||f||^2)^2$, then $\nabla C = (2(||f||^2))(2f)$

- The evaluation functional evaluates f at the specified x: $F_x[f] = f(x) = e_x[f]$
  - Its gradient is $\nabla e_x = K(x, \cdot)$

$$\begin{aligned} e_x[f + \epsilon g] &= f(x) + \epsilon g(x) + 0 \\ &= f(x) + \epsilon\langle K(x, \cdot), g\rangle + 0 \\ &= e_x[f] + \epsilon\langle \nabla e_x, g\rangle + O(\epsilon^2) \end{aligned}$$

  - Called a *linear functional* due to lack of multiplier on perturbation ϵ

4 Functional gradient descent

- Consider the regularized least squares loss function $L[f]$

$$\begin{aligned} L[f] &= (f(x_i) - y_i)^2 + \lambda ||f||^2 \\ \nabla L[f] &= 2(f(x_i) - y_i)K(x_i, \cdot) + 2\lambda f \end{aligned}$$

- Update rule:

$$\begin{aligned} f^{t+1} &\leftarrow f^t - \eta_t \nabla L \\ &\leftarrow f^t - \eta_t(2(f^t(x_i) - y_i)K(x_i, \cdot) + 2\lambda f^t) \\ &\leftarrow f^t(1 - 2\eta_t\lambda) - \eta_t(2(f^t(x_i) - y_i)K(x_i, \cdot)) \end{aligned}$$

- Need to perform $O(T)$ work at each time step
- Example: Figure 4 shows an update over 3 points $\{(x_1, +), (x_2, -), (x_3, +)\}$. The individual kernels centered at the points are **independently** drawn with colored lines. After 3 updates, the function f looks like the solid black line.

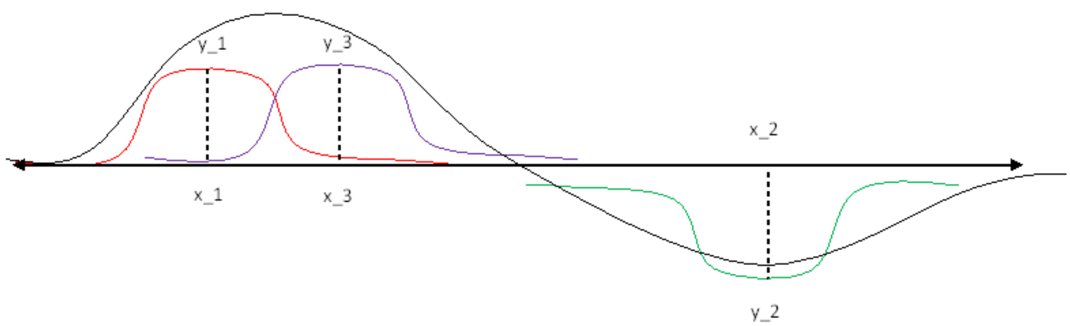

Figure 1: Illustration of function after 3 updates

- **Representer Theorem** (informally): Given a loss function and regularizer objective with many data points $\{x_i\}$, the minimizing solution f^* can be represented as

$$f^*(\cdot) = \sum_i \alpha_i K(x_i, \cdot) \tag{3}$$

- Alternate idea from class: perform gradient descent in the space of α coefficients: $\nabla_\alpha L$
  - Takes n^2 iterations to get same performance (n = number of iterations of functional gradient descent)
  - Every iteration is $O(T^2)$

5 Kernel SVM

- General loss function: $L[f] = \frac{\lambda}{2}||f||^2 + C_t(F_{x_i}[f])$
- General update rule: $f_{t+1} \leftarrow f_t(1 - \lambda\eta_t) - \eta_t C_t'(F_{x_i}[f])K(x_i, \cdot)$
- SVM cost function: $C_t(F_{x_i}) = \max(0, 1 - f(x_i)y_i)$

$$\nabla C_t = \begin{cases} 0 & , 1 - y_i f(x_i) \leq 0 \\ (C'(F_{x_i}[f]))(\nabla F_{x_i}[f]) = (-y_i)(K(x_i, \cdot)) & , \text{otherwise} \end{cases} \tag{4}$$