Kernels

Gautam Kamath

Motivation

- Most methods we've looked at so far are linear, which is limiting
- (Draw classifier with circle boundary)
- Consider the *feature map* $(x_1, x_2) \rightarrow (x_1^2, x_2^2)$
- (Draw points after the mapping)
- Data which is not linearly separable becomes linearly separable in the new space
- (Draw XOR example)
- Consider feature map $\phi: \mathbf{R}^2 \rightarrow \mathbf{R}^3$ where $\phi(x) = [x_1, x_2, x_1 x_2]$
  - Then $\text{sign}(x_1 x_2)$ works
- (Fail at drawing this in 3D)

Feature Maps

- A feature map ϕ maps a feature vector x to some higher-dimensional space
- Simple example: the padding trick
- $\phi(x) = [x, 1]$ and parameter vector $w = [p, b]$, both $\in \mathbf{R}^{d+1}$
- Before: classifiers like $\langle x, p \rangle > 0$
- After: classifiers like $\langle \phi(x), w \rangle = \langle x, p \rangle + b > 0$

Quadratic Feature Maps

- Instead of functions $x^T p + b > 0$, consider $x^T Q x + \sqrt{2} x^T p + b > 0$
  - $Q \in \mathbf{R}^{d \times d}, p \in \mathbf{R}^d, b \in \mathbf{R}$
  - Trust me on the $\sqrt{2}$ for now
- Note that $x^T Q x = \sum (x_i x_j) Q_{ij}$
- A dot product between "flattenings" of xx^T and Q, name $\overrightarrow{xx^T}$ and $\vec{Q}$ (draw)
- Take $\phi(x) = \left[\overrightarrow{xx^T}, \sqrt{2}x, 1\right]$ and $w = [\vec{Q}, p, b]$, both in $\mathbf{R}^{d^2+d+1}$
- Then $\langle \phi(x), w \rangle \leftrightarrow x^T Q x + \sqrt{2} x^T p + b$

Feature Maps

- Generally: instead of taking dot product of feature vector with parameter vector, map feature vector and take dot product with new parameter vector
- With quadratic feature map $\phi: \mathbf{R}^d \rightarrow \mathbf{R}^{d^2+d+1}$, computations go from $O(d)$ to $O(d^2)$
- What if we map to a very high-dimensional space?
- What if we have an infinite-dimensional feature map $\phi: \mathbf{R}^d \rightarrow \mathbf{R}^\infty$?
  - E.g., $(d = 1)\ \phi(x) = [1, x, x^2, x^3, x^4, \ldots]$
  - Naively, can't be computed in finite time. What do?

Using Feature Maps in SVM (Dual)

$$\min_{\alpha \in \mathbf{R}^n, C \geq \alpha \geq 0} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_i \alpha_i \text{ s.t.} \sum_i \alpha_i y_i = 0$$

Note: going back to x_i being ith point (as in previous lectures) rather than ith dimension

Using Feature Maps in SVM (Dual)

$$\min_{\alpha \in \mathbf{R}^n, C \geq \alpha \geq 0} \sum \sum \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle - \sum_i \alpha^{(i)} \text{ s.t.} \sum_i \alpha^{(i)} y^{(i)} = 0$$

- Focus on quadratic $\phi(x) = \left[\overrightarrow{xx^T}, \sqrt{2}x, 1\right]$ for now
- How to compute $\langle \phi(x), \phi(x') \rangle$?
- Naively: compute $\phi(x)$, compute $\phi(x')$, take their dot product. $O(d^2)$ time
- BUT! We don't care about $\phi(x)$ or $\phi(x')$, only their dot product $\langle \phi(x), \phi(x') \rangle$

Quadratic Feature Map Kernel

- How to compute $\langle \phi(x), \phi(x') \rangle$?

$$\langle \phi(x), \phi(x') \rangle = \left\langle \overrightarrow{xx^T}, \overrightarrow{x'x'^T} \right\rangle + \langle \sqrt{2}x, \sqrt{2}x' \rangle + \langle 1,1 \rangle$$

$$\left\langle \overrightarrow{xx^T}, \overrightarrow{x'x'^T} \right\rangle = \sum_{i,j} x_i x_j x'_i x'_j = \sum_{i,j} x_i x'_i x_j x'_j = \sum_i x_i x'_i \left(\sum_j x_j x'_j \right) = \langle x, x' \rangle^2$$

$$\langle \phi(x), \phi(x') \rangle = \langle x, x' \rangle^2 + 2\langle x, x' \rangle + 1 = (\langle x, x' \rangle + 1)^2$$

- Note that $(\langle x, x' \rangle + 1)^2$ can be computed in $O(d)$ time, instead of $O(d^2)$!
- $k(x, x') = (\langle x, x' \rangle + 1)^2$ is a *kernel*

Kernels

- $k: \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$ is a *kernel* if there exists a feature map $\phi: \mathbf{R}^d \to \mathbf{R}^m$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$
- $\phi(x)$ may be expensive (or impossible) to compute, but the kernel k may be tractable
  - Compare $O(d^2)$ for quadratic feature map versus $O(d)$ for kernel
- Polynomial kernel of degree t: $k(x, x') = (\langle x, x' \rangle + 1)^t$
- Gaussian/radial basis function: $k(x, x') = \exp(\|x - x'\|_2^2)$

What makes a valid Kernel?

- First off, if you can construct a corresponding feature map ϕ
- This also implies the following alternate interpretation
- Let $x_1, \ldots, x_n$ be an arbitrary dataset
- Let $K \in \mathbf{R}^{n \times n}$ be a matrix where $K_{ij} = k(x_i, x_j)$
- K is symmetric ($K_{ij} = K_{ji}$) and positive semidefinite
  - $v^T K v \geq 0$ for all vectors $v \in \mathbf{R}^n$
- The existence of a feature map implies these properties due to the Gram matrix (draw)

Using Kernels in SVM (Dual)

Solve

$$\min_{\alpha \in \mathbf{R}^n, C \geq \alpha \geq 0} \sum \sum \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} K_{ij} - \sum_i \alpha^{(i)} \text{ s.t.} \sum_i \alpha^{(i)} y^{(i)} = 0$$

- $K_{ij} = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle = k(x^{(i)}, x^{(j)})$
- How to classify new point?
  - $w = \sum \alpha^{(i)} y^{(i)} \phi(x^{(i)})$, but can't compute $\phi(x^i)$
  - $\text{sign}(\langle w, \phi(x) \rangle) = \text{sign}(\langle \sum \alpha^{(i)} y^{(i)} \phi(x^{(i)}), \phi(x) \rangle) = \text{sign}(\sum \alpha^{(i)} y^{(i)} k(x^{(i)}, x))$
- SVM (Linear Kernel): $O(nd)$ train time, $O(d)$ test time
- General Kernel: $O(n^2 d)$ train time, $O(nd)$ test time