The Kernel Trick for Distances

Leading discussion

Branislav Kveton
Intelligent Systems Program

Content

- Pattern classification and kernels
- Dot product as a kernel
- The kernel trick
- Positive definite (reproducing) kernels
- Feature map for PD kernels
- What is wrong with dot product?
- Dot product and squared distance
- Conditionally positive definite kernels
- Squared distance and CPD kernels
- Symmetric kernels

Pattern Classification and Kernels

- let us have some training data of m elements

$$(x_1, y_1), \dots, (x_m, y_m) \in X \times Y$$

- where X is a set of patterns and Y is a set of classifications
- to classify an unseen pattern x, one takes into account a notion of similarity between already classified x_is and x
- the similarity measure can be formalized as

$$k: \quad X \times X \to R, \quad (x, x') \mapsto k(x, x')$$

- and k is called a kernel
- further derivations assume real-value symmetric kernels

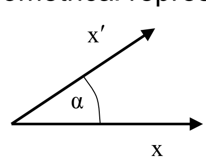
$$k(x, x') = k(x', x)$$

Dot Product as a Kernel

- is a similarity measure of the form

$$\langle x, x' \rangle = \sum_{i=1}^{n} x_i x_i'$$

- geometrical representation



$$\langle x, x' \rangle = \|x\| \|x'\| \cos \alpha$$

$$\langle x, x' \rangle = \begin{cases} 0 & x \perp x' \\ \|x\| \|x'\| & x \parallel x' \\ (0, \|x\| \|x'\|) & \text{else} \end{cases}$$

- is one of the simplest kernels

The Kernel Trick

- before a learning algorithm is used, input space X is usually mapped into a feature space F by transformation $\varphi: X \to F$
- to avoid the computation in a potentially high dimensional space F, one picks features such that the dot product in the feature space can be evaluated by a non-linear function in the input space, known as the kernel trick

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle$$

Positive Definite (Reproducing) Kernels

- gram matrix K with respect to $x_1, \ldots, x_m$ is defined as

$$K_{i,j} = k(x_i, x_j)$$

- gram matrix for the dot kernel with respect to x_1 and x_2 is

$$\begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_2, x_1 \rangle \\ \langle x_1, x_2 \rangle & \langle x_2, x_2 \rangle \end{pmatrix}$$

- a real symmetric matrix K is positive definite if for every c

$$cKc^T = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j K_{i,j} \geq 0$$

- a kernel is positive definite (PD) if the corresponding gram matrix is positive definite. In such a case, there exists a procedure to construct the feature space associated with φ

Feature Map for PD Kernels

- define a feature map

$$\varphi : X \mapsto R^X, \quad x \mapsto k(.,x)$$

- form a linear combination of basis functions

$$f(.) = \sum_{i=1}^{m} \alpha_i k(.,x_i), \quad g(.) = \sum_{j=1}^{m'} \beta_j k(.,x'_j)$$

- define the following operator

$$\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j), \quad k(x,x') = \langle k(.,x), k(.,x') \rangle = \langle \varphi(x), \varphi(x') \rangle$$

- and prove that
  - the operator is in fact dot product
  - the operation is a PD kernel

What is Wrong with Dot Product?

- if patterns x and x' are translated by

$$x \mapsto x - x_0, \quad x' \mapsto x' - x_0$$

- the dot product between the pattern changes
- this is not suitable for algorithms where the learning process should be translation invariant (PCA)
- squared distance as a dissimilarity measure of the form

$$\|x - x'\|^2$$

- is translation invariant. Moreover, it can be expressed in the feature space by the kernel trick

$$\begin{aligned} \|\varphi(x) - \varphi(x')\|^2 &= \langle \varphi(x), \varphi(x) \rangle - 2\langle \varphi(x), \varphi(x') \rangle + \langle \varphi(x'), \varphi(x') \rangle \\ &= k(x,x) + k(x',x') - 2k(x,x') \end{aligned}$$

Dot Product and Squared Distance

- dot product and squared distance measures can be related in the translated space by

$$\begin{aligned}
\langle x - x_0, x' - x_0 \rangle &= \tfrac{1}{2}\left(-\|x - x'\|^2 + \|x - x_0\|^2 + \|x_0 - x'\|^2\right) \\
2\langle x - x_0, x' - x_0 \rangle &= -\|x - x'\|^2 + \|x - x_0\|^2 + \|x_0 - x'\|^2 \\
2\langle x, x' \rangle - 2\langle x, x_0 \rangle - 2\langle x', x_0 \rangle + 2\langle x_0, x_0 \rangle &= -\left(\langle x, x \rangle - 2\langle x, x' \rangle + \langle x', x' \rangle\right) + \left(\langle x, x \rangle - 2\langle x, x_0 \rangle + \langle x_0, x_0 \rangle\right) + \left(\langle x_0, x_0 \rangle - 2\langle x', x_0 \rangle + \langle x', x' \rangle\right)
\end{aligned}$$

- the dot product is a PD kernel

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \langle x_i - x_0, x_j - x_0 \rangle = \sum_{i=1}^{m} c_i (x_i - x_0)^T \sum_{j=1}^{m} c_j (x_j - x_0) \\
&= \left(\sum_{i=1}^{m} c_i (x_i - x_0)^T\right)\left(\sum_{i=1}^{m} c_i (x_i - x_0)\right) = \left\|\sum_{i=1}^{m} c_i (x_i - x_0)\right\|^2 \geq 0
\end{aligned}$$

Conditionally Positive Definite Kernels

- a kernel is conditionally positive definite (CPD) if for every c

$$cKc^T = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j K_{i,j} \geq 0, \quad \sum_{i=1}^{m} c_i = 0$$

- $k(x, x')$ is a PD kernel if and only if $q(x, x')$ is a CPD kernel

$$k(x, x') = q(x, x') - q(x, x_0) - q(x_0, x') + q(x_0, x_0)$$

- negative squared distance is a CPD kernel

$$\begin{aligned}
-\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j q(x_i, x_j) &= -\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \|x_i - x_j\|^2 \\
&= -\sum_{i=1}^{m} c_i \sum_{j=1}^{m} c_j \|x_j\|^2 - \sum_{j=1}^{m} c_j \sum_{i=1}^{m} c_i \|x_i\|^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \langle x_i, x_j \rangle \\
&= 2\left(\sum_{i=1}^{m} c_i x_i^T\right)\left(\sum_{i=1}^{m} c_i x_i\right) = 2\left\|\sum_{i=1}^{m} c_i x_i\right\|^2 \geq 0
\end{aligned}$$

Squared Distance and CPD Kernels

- implies that q(x, x') of the following form are CPD kernels

$$q(x,x') = -\|x - x'\|^{\beta}, \quad 0 < \beta \le 2$$

- CDP kernels can be used to define the squared distance measure in some feature space

$$\begin{aligned}
\|\varphi(x) - \varphi(x')\|^2 &= \langle \varphi(x), \varphi(x) \rangle - 2\langle \varphi(x), \varphi(x') \rangle + \langle \varphi(x'), \varphi(x') \rangle \\
&= k(x,x) + k(x',x') - 2k(x,x') \\
& q(x,x) - q(x,x_0) - q(x_0,x) + q(x_0,x_0) + \\
&= q(x',x') - q(x',x_0) - q(x_0,x') + q(x_0,x_0) - \\
& (2q(x,x') - 2q(x,x_0) - 2q(x_0,x') + 2q(x_0,x_0)) \\
&= -q(x,x') + \tfrac{1}{2}(q(x,x) + q(x',x'))
\end{aligned}$$

- depending on the choice of β, the squared distance measure is used in an appropriate feature space

Symmetric Kernels

- construction similar to the feature maps of PD kernels can be done for symmetric kernels

$$q(x,x') = Q(\varphi(x), \varphi(x'))$$

- as the assumption of q(x, x') being PD kernel is dropped, Q does not fulfill requirements for dot product

$$Q(f,f) = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j q(x_i, x_j) \ge 0$$

- generalization of PD-CPD proposition for symmetric kernels

$$\begin{aligned}
k(x,x') &= Q(\varphi(x) - \varphi(x_0), \varphi(x') - \varphi(x_0)) \\
&= Q(\varphi(x), \varphi(x')) - Q(\varphi(x), \varphi(x_0)) - Q(\varphi(x'), \varphi(x_0)) + Q(\varphi(x_0), \varphi(x_0)) \\
&= q(x,x') - q(x,x_0) - q(x',x_0) + q(x_0,x_0)
\end{aligned}$$

Symmetric Kernels

- a symmetric kernel $q(x, x')$ is a CPD kernel if and only if $k(x, x')$ is a PD kernel

$$k(x,x') = \tfrac{1}{2}\left(q(x,x') - \sum_{i=1}^{m} c_i q(x,x_i) - \sum_{i=1}^{m} c_i q(x_i,x') + \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j q(x_i,x_j) \right), \quad \sum_{i=1}^{m} c_i = 1$$

- this is a generalization of the previous results with respect to an arbitrary center in the space, which is weighted by c_i

Thank You for Listening