CPSC 340:
Machine Learning and Data Mining

Kernel Methods

Bonus slides

Kernel Trick for Non-Vector Data

- Consider data that doesn't look like this:

$$X = \begin{bmatrix} 0.5377 & 0.3188 & 3.5784 \\ 1.8339 & -1.3077 & 2.7694 \\ -2.2588 & -0.4336 & -1.3499 \\ 0.8622 & 0.3426 & 3.0349 \end{bmatrix}, \quad y = \begin{bmatrix} +1 \\ -1 \\ -1 \\ +1 \end{bmatrix},$$

- But instead looks like this:

$$X = \begin{bmatrix} \text{Do you want to go for a drink sometime?} \\ \text{J'achète du pain tous les jours.} \\ \text{Fais ce que tu veux.} \\ \text{There are inner products between sentences?} \end{bmatrix}, y = \begin{bmatrix} +1 \\ -1 \\ -1 \\ +1 \end{bmatrix}.$$

- We can interpret k(xi,xj) as a "similarity" between objects xi and xj.
  - We don't need features if we can compute "similarity" between objects.
  - Kernel trick lets us fit regression models without explicit features.
  - There are "string kernels", "image kernels", "graph kernels", and so on.

Kernel Trick for Non-Vector Data

- Recent list of types of data where people have defined kernels:

trees (Collins & Duffy, 2001; Kashima & Koyanagi, 2002), time series (Cuturi, 2011), strings (Lodhi et al., 2002), mixture models, hidden Markov models or linear dynamical systems (Jebara et al., 2004), sets (Haussler, 1999; Gärtner et al., 2002), fuzzy domains (Guevara et al., 2017), distributions (Hein & Bousquet, 2005; Martins et al., 2009; Muandet et al., 2011), groups (Cuturi et al., 2005) such as specific constructions on permutations (Jiao & Vert, 2016), or graphs (Vishwanathan et al., 2010; Kondor & Pan, 2016).

- Bonus slide overviews a particular “string” kernel.

https://arxiv.org/pdf/1802.04784.pdf

Valid Kernels

- What kernel functions $k(x_i,x_j)$ can we use?
- Kernel 'k' must be an inner product in some space:
  - There must exist a mapping from the x_i to some z_i such that $k(x_i,x_j) = z_i^T z_j$.
- It can be hard to show that a function satisfies this.
  - Infinite-dimensional eigenfunction problem.
- But like convex functions, there are some simple rules for constructing "valid" kernels from other valid kernels (bonus slide).

Kernel Trick for Other Methods

- Besides L2-regularized least squares, when can we use kernels?
  - We can compute Euclidean distance with kernels:

$$\|z_i - z_j\|^2 = z_i^T z_i - 2 z_i^T z_j + z_j^T z_j = k(x_i, x_i) - 2k(x_i, x_j) + k(x_j, x_j)$$

  - All of our distance-based methods have kernel versions:
    - Kernel k-nearest neighbours.
    - Kernel clustering k-means (allows non-convex clusters)
    - Kernel density-based clustering.
    - Kernel hierarchical clustering.
    - Kernel distance-based outlier detection.
    - Kernel “Amazon Product Recommendation”.

Kernel Trick for Other Methods

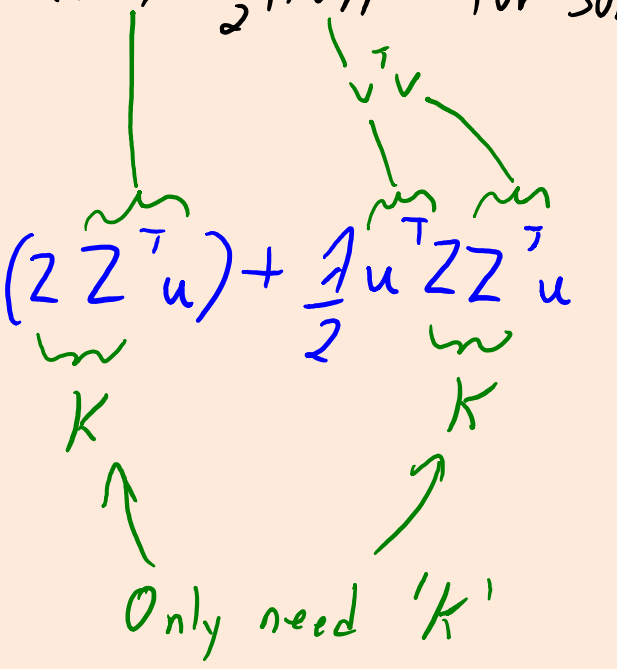
- Besides L2-regularized least squares, when can we use kernels?
  - "Representer theorems" (bonus slide) have shown that any L2-regularized linear model can be kernelized:

If learning can be written in the form $\min_v f(Zv) + \frac{1}{2}\|v\|^2$ for some 'Z'

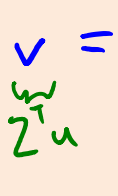
then under weak conditions ("representer theorem")

we can re-parameterize in terms of $v = Z^T u$

giving $\min_u f(ZZ^T u) + \frac{1}{2} u^T Z Z^T u$



At test time you would use $\tilde{Z}v = \tilde{Z}Z^T u = \tilde{K}u$



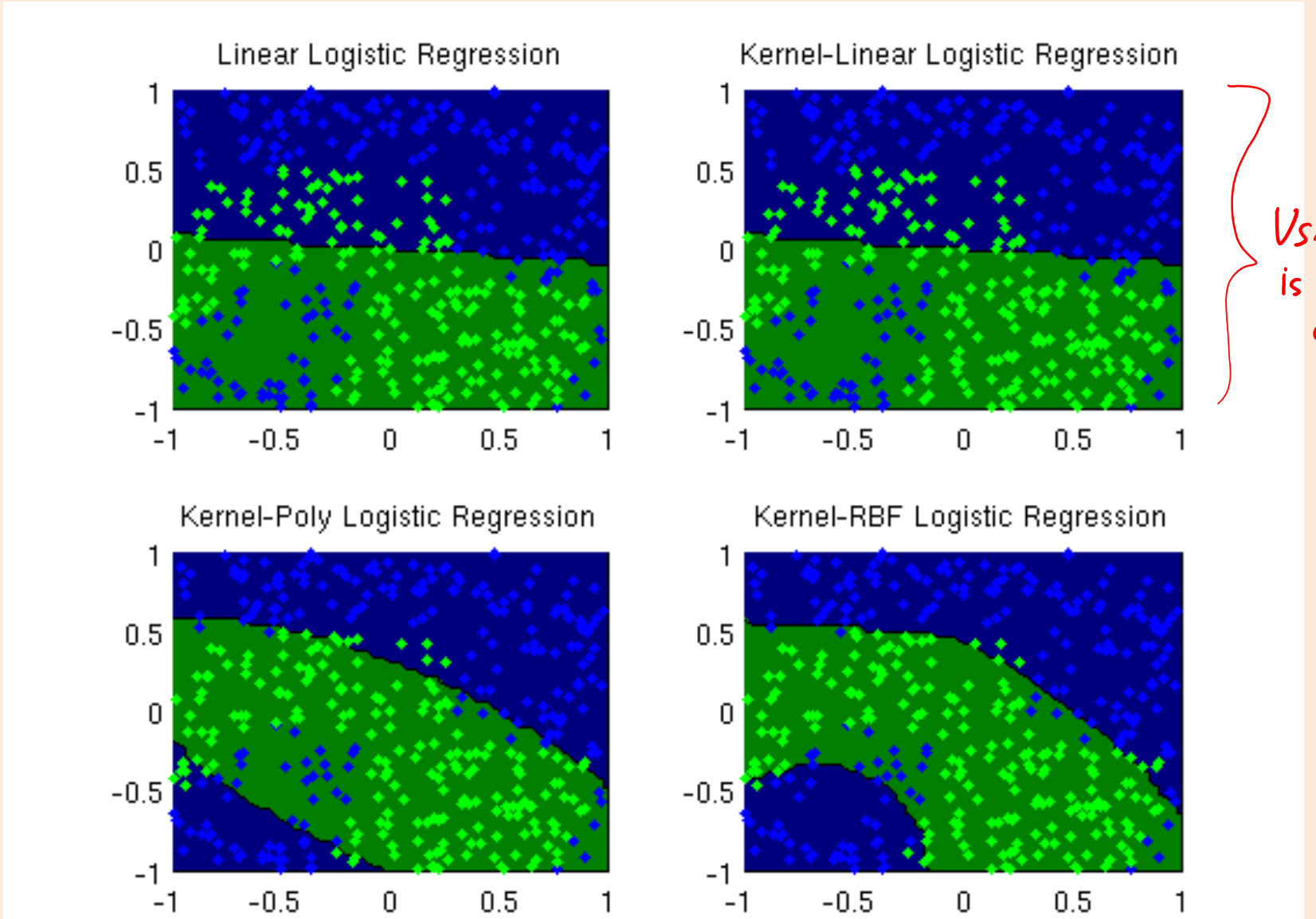
Kernel Trick for Other Methods

- Besides L2-regularized least squares, when can we use kernels?
  - "Representer theorems" (bonus slide) have shown that any L2-regularized linear model can be kernelized:
    - L2-regularized robust regression.
    - L2-regularized brittle regression.
    - L2-regularized logistic regression.
    - L2-regularized hinge loss (SVMs).

With a particular implementation, can reduce prediction cost from $O(ndt)$ to $O(mdt)$.

m: Number of support vectors.

Logistic Regression with Kernels



Using "linear" Kernel is the same as using original features

Feature Selection Hierarchy

- Consider a linear models with higher-order terms,

$$\hat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + w_{12} x_{i1} x_{i2} + w_{13} x_{i1} x_{i3} + w_{23} x_{i2} x_{i3} + w_{123} x_{i1} x_{i2} x_{i3}$$

- The number of higher-order terms may be too large.
  - Can't even compute them all.
  - We need to somehow decide which terms we'll even consider.

- Consider the following hierarchical constraint:
  - You only allow $w_{12} \neq 0$ if $w_1 \neq 0$ and $w_2 \neq 0$.
  - "Only consider feature interaction if you are using both features already."

Hierarchical Forward Selection

- Hierarchical Forward Selection:
  - Usual forward selection, but consider interaction terms obeying hierarchy.
  - Only consider $w_{12} \neq 0$ once $w_1 \neq 0$ and $w_2 \neq 0$.
  - Only allow $w_{123} \neq 0$ once $w_{12} \neq 0$ and $w_{13} \neq 0$ and $w_{23} \neq 0$.
  - Only allow $w_{1234} \neq 0$ once all threeway interactions are present.

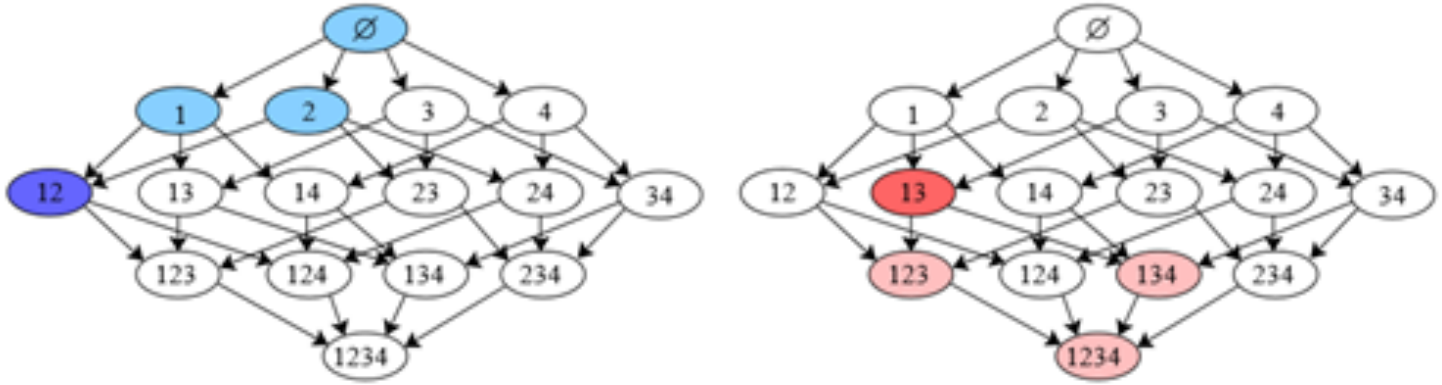


Fig 9: Power set of the set $\{1, \ldots, 4\}$: in blue, an authorized set of selected subsets. In red, an example of a group used within the norm (a subset and all of its descendants in the DAG).

http://arxiv.org/pdf/1109.2397v2.pdf

Bonus Slide: Equivalent Form of Ridge Regression

Note that $\hat{X}$ and Y are the same on the left and right side, so we only need to show that

$$(X^T X + \lambda I)^{-1} X^T = X^T (X X^T + \lambda I)^{-1}. \tag{1}$$

A version of the matrix inversion lemma (Equation 4.107 in MLAPP) is

$$(E - FH^{-1}G)^{-1} FH^{-1} = E^{-1}F(H - GE^{-1}F)^{-1}.$$

Since matrix addition is commutative and multiplying by the identity matrix does nothing, we can re-write the left side of (1) as

$$(X^T X+\lambda I)^{-1} X^T = (\lambda I+X^T X)^{-1} X^T = (\lambda I+X^T I X)^{-1} X^T = (\lambda I-X^T(-I)X)^{-1} X^T = -(\lambda I-X^T(-I)X)^{-1} X^T(-I)$$

Now apply the matrix inversion with $E = \lambda I$ (so $E^{-1} = \left(\frac{1}{\lambda}\right) I$), $F = X^T$, $H = -I$ (so $H^{-1} = -I$ too), and $G = X$:

$$-(\lambda I - X^T(-I)X)^{-1} X^T(-I) = -(\frac{1}{\lambda}) I X^T(-I - X\left(\frac{1}{\lambda}\right) X^T)^{-1}.$$

Now use that $(1/\alpha)A^{-1} = (\alpha A)^{-1}$, to push the $(-1/\lambda)$ inside the sum as $-\lambda$,

$$-(\frac{1}{\lambda}) I X^T(-I - X\left(\frac{1}{\lambda}\right) X^T)^{-1} = X^T(\lambda I + X X^T)^{-1} = X^T(X X^T + \lambda I)^{-1}.$$

Why is inner product a similarity?

- It seems weird to think of the inner-product as a similarity.
- But consider this decomposition of squared Euclidean distance:

$$\frac{1}{2}\|x_i - x_j\|^2 = \frac{1}{2}\|x_i\|^2 - x_i^T x_j + \frac{1}{2}\|x_j\|^2$$

- If all training examples have the same norm, then minimizing Euclidean distance is equivalent to maximizing inner product.
  - So "high similarity" according to inner product is like "small Euclidean distance".
  - The only difference is that the inner product is biased by the norms of the training examples.
  - Some people explicitly normalize the x_i by setting $x_i = (1/||x_i||)x_i$, so that inner products act like the negation of Euclidean distances.
    - E.g., Amazon product recommendation.

Guasian-RBF Kernels

- The most common kernel is the Gaussian-RBF (or 'squared exponential') kernel,

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right).$$

- What function $\phi(x)$ would lead to this as the inner-product?
  - To simplify, assume $d = 1$ and $\sigma = 1$,

$$\begin{aligned} k(x_i, x_j) &= \exp(-x_i^2 + 2x_i x_j - x_j^2) \\ &= \exp(-x_i^2)\exp(2x_i x_j)\exp(-x_j^2), \end{aligned}$$

    so we need $\phi(x_i) = \exp(-x_i^2)z_i$ where $z_i z_j = \exp(2x_i x_j)$.
    - For this to work for *all* x_i and x_j, z_i must be infinite-dimensional.
  - If we use that

$$\exp(2x_i x_j) = \sum_{k=0}^{\infty} \frac{2^k x_i^k x_j^k}{k!},$$

    then we obtain

$$\phi(x_i) = \exp(-x_i^2)\left[1 \quad \sqrt{\tfrac{2}{1!}}x_i \quad \sqrt{\tfrac{2^2}{2!}}x_i^2 \quad \sqrt{\tfrac{2^3}{3!}}x_i^3 \quad \cdots\right].$$

question

stop following 83 views

Why RBF-kernel not the same as RBF-basis?

I do not quite understand the two statements in red box? I think with k as defined that way, it is just the $g(||x_i - x_j||)$ as we saw in the last lecture of RBF basis? Why they are not equivalent? What does "equivalent" here mean?

Also, why now "we are using them as inner product"? Is it because we now regard $k(x_i, x_j)$ as the inner product of z_i and z_j, which are some magical transformation of x_i and x_j? (Like $k(x_i, x_j) = (1 + x_i^T x_j)^p$ is the inner product of z_i and z_j, which are polynomial transformation of x_i and x_j)?

Chenliang Zhou 8 months ago Oh so is my following reasoning correct?:

Let Z and $\tilde{Z}$ be as defined in lecture 22a.

In Gaussian RBF basis, $\tilde{y} = \tilde{Z}(Z^T Z + \lambda I)^{-1} Z^T y = \tilde{Z} Z^T (Z Z^T + \lambda I)^{-1} y$.

In Gaussian RBF kernel, we have $\tilde{y} = \tilde{K}(K + \lambda I)^{-1} y$ where where K and $\tilde{K}$ are those 2 horrible matrices for Gaussian RBF kernels. Since they are the same formula, $K = Z$ and $\tilde{K} = \tilde{Z}$, so $\tilde{y} = \tilde{Z}(Z + \lambda I)^{-1} y$.

So Gaussian RBF basis and Gaussian RBF kernel are different because in general,
$\tilde{Z} Z^T (Z Z^T + \lambda I)^{-1} (\text{for G-RBF basis}) \neq \tilde{Z}(Z + \lambda I)^{-1} (\text{for G-RBF kernel})$.

A String Kernel

- A classic “string kernel”:
  - We want to compute k(“cat”, “cart”).
  - Find all common subsequences: ‘c’, ‘a’, ‘t’, ‘ca’, ‘at’, ‘ct’, ‘cat’.
  - Weight them by total length in original strings:
    - ‘c’ has length (1,1), ‘ca’ has lengths (2,2), ‘ct’ has lengths (3,4), and so on.
  - Add up the weighted lengths of common subsequences to get a similarity:

$$k(\text{“cat”}, \text{“cart”}) = \underbrace{\gamma^1\gamma^1}_{\text{‘c’}} + \underbrace{\gamma^1\gamma^1}_{\text{‘a’}} + \underbrace{\gamma^1\gamma^1}_{\text{‘t’}} + \underbrace{\gamma^2\gamma^2}_{\text{‘ca’}} + \underbrace{\gamma^2\gamma^3}_{\text{‘at’}} + \underbrace{\gamma^3\gamma^4}_{\text{‘ct’}} + \underbrace{\gamma^3\gamma^4}_{\text{‘cat’}},$$

  where γ is a hyper-parameter controlling influence of length.

- Corresponds to exponential feature set (counts/lengths of all subsequences).
  - But kernel can be computed in polynomial time by dynamic programming.
- Many variations exist.

Constructing Valid Kernels

- If $k_1(x_i, x_j)$ and $k_2(x_i, x_j)$ are valid kernels, then the following are valid kernels:
  - $k_1(\phi(x_i), \phi(x_j))$.
  - $\alpha k_1(x_i, x_j) + \beta k_2(x_i, x_j)$ for $\alpha \geq 0$ and $\beta \geq 0$.
  - $k_1(x_i, x_j) k_2(x_i, x_j)$.
  - $\phi(x_i) k_1(x_i, x_j) \phi(x_j)$.
  - $\exp(k_1(x_i, x_j))$.
- Example: Gaussian-RBF kernel:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

$$= \underbrace{\exp\left(-\frac{\|x_i\|^2}{\sigma^2}\right)}_{\phi(x_i)} \underbrace{\exp\left(\underbrace{\frac{2}{\sigma^2}}_{\alpha \geq 0} \underbrace{x_i^T x_j}_{\text{valid}}\right)}_{\exp(\text{valid})} \underbrace{\exp\left(-\frac{\|x_j\|^2}{\sigma^2}\right)}_{\phi(x_j)}.$$

Representer Theorem

- Consider linear model differentiable with losses f_i and L2-regularization,

$$\underset{w \in \mathbb{R}^d}{\text{argmin}} \sum_{i=1}^{n} f_i(w^T x_i) + \frac{\lambda}{2}\|w\|^2.$$

- Setting the gradient equal to zero we get

$$0 = \sum_{i=1}^{n} f_i'(w^T x_i) x_i + \lambda w.$$

- So any solution w^* can written as a linear combination of features x_i,

$$\begin{aligned} w^* &= -\frac{1}{\lambda} \sum_{i=1}^{n} f_i'((w^*)^T x_i) x_i = \sum_{i=1}^{n} z_i x_i \\ &= X^T z. \end{aligned}$$

- This is called a representer theorem (true under much more general conditions).

Kernel Trick for Other Methods

- Besides L2-regularized least squares, when can we use kernels?
  - "Representer theorems" have shown that
    any L2-regularized linear model can be kernelized.

  - Linear models without regularization fit with gradient descent.
    - If you starting at v=0 or with any other value in span of rows of 'Z'.

Iterations of gradient descent on $f(Zv)$ can be written as $v = Z^T u$
which lets us re-parameterize as $f(ZZ^T u)$

At test time you would use $\tilde{Z}v = \tilde{Z}Z^T u = \tilde{K}u$

($X^T u$ under v; $\tilde{K}$ under $\tilde{Z}Z^T$)