

Assignment Project Exam Help

Machine learning lecture slides

COMS 4771 Fall 2020

<https://powcoder.com>

Add WeChat powcoder

Regression III: Kernels

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- ▶ Dual form of ridge regression
- ▶ Examples of kernel trick
- ▶ Kernel methods

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- Let $A = \frac{1}{\sqrt{n}} \begin{bmatrix} \leftarrow & x_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & x_n^\top & \rightarrow \end{bmatrix} \in \mathbb{R}^{n \times d}$ and $b = \frac{1}{\sqrt{n}} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$
- Linear algebraic identity: for any $A \in \mathbb{R}^{n \times d}$ and any $\lambda > 0$,

$$\underbrace{(A^\top A + \lambda I)^{-1}}_{d \times d} A^\top = A^\top \underbrace{(A A^\top + \lambda I)^{-1}}_{n \times n}.$$

- Check: multiply both sides by $A^\top A + \lambda I$ and “factor”.

Add WeChat powcoder

Alternative (dual) form for ridge regression (1)

- Implications for ridge regression

$$w = \underbrace{A^T(AA^T + \lambda I)^{-1}b}_{=:\sqrt{n}\hat{\alpha}} = \sqrt{n}A^T\hat{\alpha} = \sum_{i=1}^n \hat{\alpha}_i x_i.$$

- Matrix $AA^T = \frac{1}{n}K$ where $K \in \mathbb{R}^{n \times n}$ is the [Gram matrix](https://powcoder.com)

$$K_{i,j} = x_i^T x_j.$$

- Prediction with w on new point x :

$$x^T \hat{w} = \sum_{i=1}^n \hat{\alpha}_i \cdot x^T x_i$$

Alternative (dual) form for ridge regression (2)

- ▶ Therefore, can “represent” predictor via data points x_1, \dots, x_n and $\hat{\alpha}$.
- ▶ Similar to nearest neighbor classifier, except also have $\hat{\alpha}$
- ▶ To get $\hat{\alpha}$: solve linear system involving K (and not A directly)
 - ▶ To make prediction on x : iterate through the x_i to compute inner products with x ; take appropriate weighted sum of results
- ▶ When is this a good idea?

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Quadratic expansion

- Suppose we want to do feature expansion to get all quadratic terms in $\varphi(x)$

$$\varphi(x) = (1, \underbrace{\sqrt{2}x_1, \dots, \sqrt{2}x_d}_{\text{linear terms}}, \underbrace{x_1^2, \dots, x_d^2}_{\text{squared terms}}, \underbrace{\sqrt{2}x_1x_2, \dots, \sqrt{2}x_1x_d, \dots}_{\text{cross terms}}, \dots)$$

- This feature expansion has $1 + 2d + \binom{d}{2} = \Theta(d^2)$ terms
 - Explicitly computing $\varphi(x)$, $\varphi(x')$, and then $\varphi(x)^\top \varphi(x')$ would take $\Theta(d^2)$ time.

Add WeChat powcoder

- “Kernel trick”: can compute $\varphi(x)^\top \varphi(x')$ in $O(d)$ time:

$$\varphi(x)^\top \varphi(x') = (1 + x^\top x')^2.$$

- Similar trick for cubic expansion, quartic expansion, etc.

- ▶ For any $\sigma > 0$, there is an infinite-dimensional feature expansion $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}^\infty$ such that

Assignment Project Exam Help

$$\varphi(x)^\top \varphi(x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right),$$

which can be computed in $O(d)$ time.

- ▶ Called Gaussian kernel or Radial Basis Function (RBF) kernel (with bandwidth σ).

Add WeChat powcoder

- ▶ Feature expansion for $d = 1$ and $\sigma = 1$ case:

$$\varphi(x) = e^{-x^2/2} \left(1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \dots\right).$$

- ▶ A positive definite kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a symmetric function satisfying the following property: For any n , and any $x_1, \dots, x_n \in \mathcal{X}$, the $n \times n$ matrix whose (i, j) th entry is $k(x_i, x_j)$ is positive semidefinite.

- ▶ Theorem: For any positive definite kernel k , there exists a feature map $\varphi: \mathcal{X} \rightarrow H$ such that $\varphi(x)^\top \varphi(x') = k(x, x')$ for all $x, x' \in \mathcal{X}$.

▶ Here, H is a special kind of inner product space called the Reproducing Kernel Hilbert Space (RKHS) corresponding to k .

- ▶ Algorithmically, we don't have to worry about what φ is. Instead, just use k .

Kernel ridge regression (1)

- ▶ Training data $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$
- ▶ Ridge regression with feature map φ : minimize

Assignment Project Exam Help

$$\frac{1}{n} \sum_{i=1}^n (\varphi(x_i)^\top w - y_i)^2 + \lambda \|w\|_2^2$$

- ▶ Compute the $n \times n$ [kernel matrix](https://powecoder.com) K where

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

- ▶ Letting $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ for $\alpha = (\alpha_1, \dots, \alpha_n)$, ridge regression objective is equivalent to

$$\frac{1}{n} \|K\alpha - y\|_2^2 + \lambda \alpha^\top K \alpha$$

where $y = (y_1, \dots, y_n) \in \mathbb{R}^n$.

Kernel ridge regression (2)

- ▶ Minimizer wrt α is solution $\hat{\alpha}$ to linear system of equations

Assignment Project Exam Help

- ▶ Return predictor that is represented by $\hat{\alpha} \in \mathbb{R}^n$ and x_1, \dots, x_n
 - ▶ To make prediction on new $x \in \mathcal{X}$: output

<https://powcoder.com>

$$\sum_{i=1}^n \hat{\alpha}_i \cdot k(x, x_i).$$

- ▶ Inductive bias:

Add WeChat powcoder

$$\begin{aligned} |\hat{w}^\top \varphi(x) - \hat{w}^\top \varphi(x')| &\leq \|\hat{w}\|_2 \cdot \|\varphi(x) - \varphi(x')\|_2 \\ &= \sqrt{\hat{\alpha}^\top K \hat{\alpha}} \cdot \|\varphi(x) - \varphi(x')\|_2 \end{aligned}$$

- ▶ Many methods / algorithms can be “kernelized” into [kernel methods](#)
 - ▶ E.g.: nearest neighbor, PCA, SVM, gradient descent . .
- ▶ “Spectral regularization” with kernels: solve $g(K/n)\alpha = y/n$ for α .

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Assignment Project Exam Help
<https://powcoder.com>

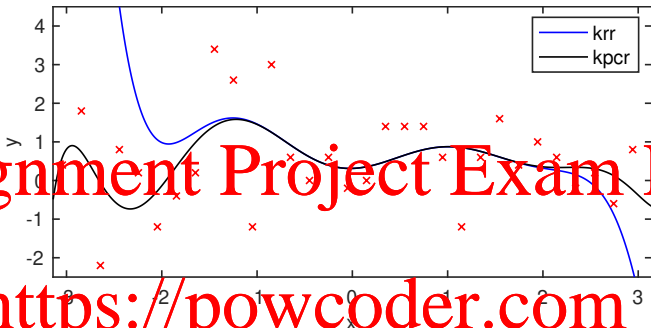
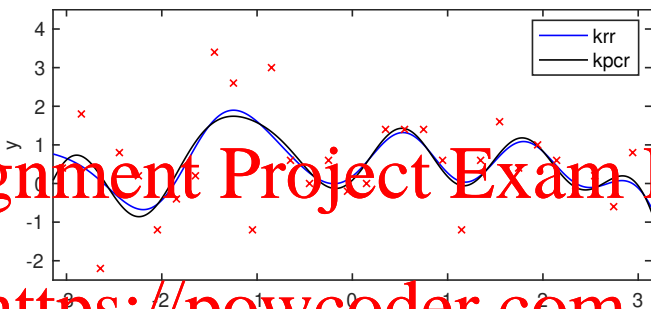


Figure 1: Polynomial kernel with Kernel Ridge Regression and Kernel PCR

Add WeChat powcoder



Assignment Project Exam Help

<https://powcoder.com>

Figure 2: RBF kernel with Kernel Ridge Regression and Kernel PCR

Add WeChat powcoder

Assignment Project Exam Help

<https://powcoder.com>

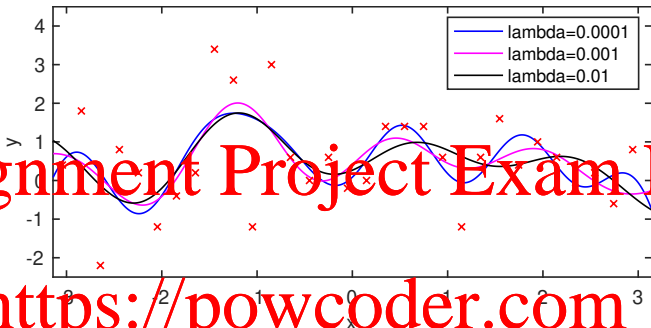


Figure 3: RBF kernel with Kernel PCR

Add WeChat powcoder

New kernels from old kernels

- ▶ Suppose k_1 and k_2 are positive definite kernel functions.
- ▶ Is $k(x, x') = k_1(x, x') + k_2(x, x')$ a positive definite kernel function?
- ▶ Is $k(x, x') = a k_1(x, x')$ (for $a \geq 0$) a positive definite kernel function?
- ▶ Is $k(x, x') = k_1(x, x') k_2(x, x')$ a positive definite kernel function?

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

- ▶ Problem with kernel methods when n is large
 - ▶ Kernel matrix K is of size n^2
 - ▶ Time for prediction generally $\propto n$
- ▶ Some possible solutions:
 - ▶ Nystrom approximations
 - ▶ Find other ways to make $\hat{\alpha}$ sparse
 - ▶ Random Fourier features

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder