

Assignment Project Exam Help

Machine learning lecture slides

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Regression III: Kernels

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

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

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Linear algebraic identity

- ▶ 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 $\lambda > 0$,

<https://eduassistpro.github.io>

- ▶ Check: multiply both sides by

Add WeChat edu_assist_pr

Alternative (dual) form for ridge regression (1)

- Implications for ridge regression

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

- <https://eduassistpro.github.io>

- Prediction with \hat{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

Assignment Project Exam Help

- ▶ <https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Quadratic expansion

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

Assignment Project Exam Help

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

- <https://eduassistpro.github.io> would take $\Theta(d^2)$ time.

Add WeChat edu_assist_pr

- “Kernel trick”: can compute $\varphi(x)$

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

- 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

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

- ▶ <https://eduassistpro.github.io/>
(with bandwidth σ).

Add WeChat edu_assist_pr

- ▶ Feature expansion for $d = 1$ and

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

- ▶ 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.

- ▶ <https://eduassistpro.github.io>) for all $x, x' \in \mathcal{X}$.

▶ Here, H is a special kind of inner product
Reproducing Kernel Hilbert Space (RKHS) 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

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

- ▶ <https://eduassistpro.github.io>

$$K_{i,j} = \langle \varphi(x_i), \varphi(x_j) \rangle$$

- ▶ Letting $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ for α , the 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

<https://eduassistpro.github.io>

$i=1$

- ▶ Inductive bias:

Add WeChat edu_assist_pro

$$|\hat{w}^\top \varphi(x) - \hat{w}^\top \varphi(x')| \leq \|\hat{w}\|$$

$$= \sqrt{\hat{\alpha}^\top K \hat{\alpha}} \cdot \|\varphi(x) - \varphi(x')\|_2$$

- ▶ Many methods / algorithms can be “kernelized” into kernel methods

Assignment Project Exam Help

- ▶ E.g.: nearest neighbor, PCA, SVM, gradient descent . . .
- ▶ “Spectral regularization” with kernels: solve $g(K/n)\alpha = y/n$

<https://eduassistpro.github.io>

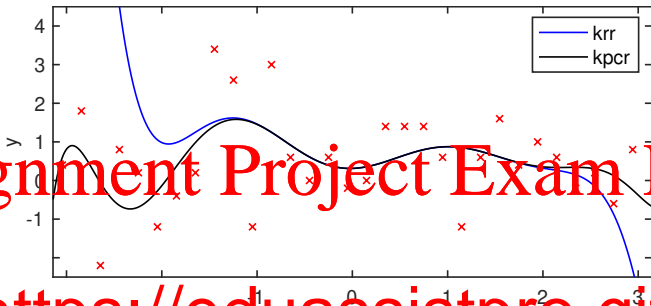
Add WeChat edu_assist_pr

Assignment Project Exam Help

<https://eduassistpro.github.io>

Figure 1: Polynomial kernel with Kernel Ridge Regress

Add WeChat edu_assist_pr

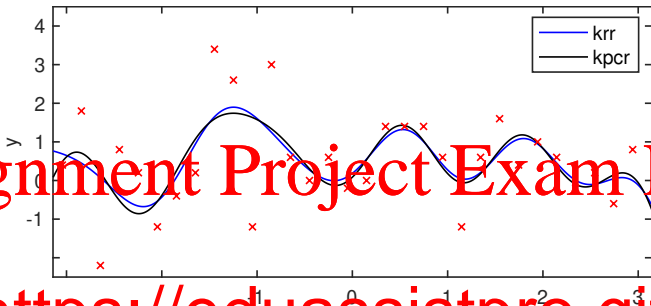


Assignment Project Exam Help

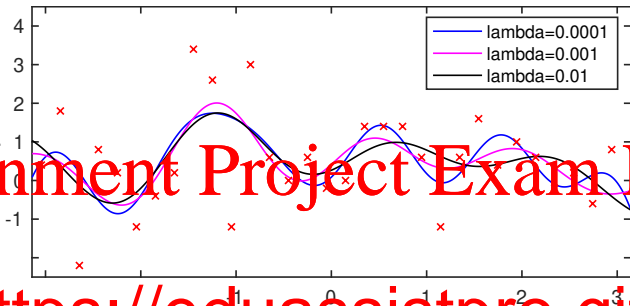
<https://eduassistpro.github.io>

Figure 2: RBF kernel with Kernel Ridge Regression

Add WeChat edu_assist_pr



Assignment Project Exam Help



<https://eduassistpro.github.io>

Figure 3: RBF kernel with Kernel

Add WeChat edu_assist_pr

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

Assignment Project Exam Help

- ▶ <https://eduassistpro.github.io>

Add WeChat edu_assist_pr

- ▶ 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:

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr