

Assignment Project Exam Help

Machine learning lecture slides

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Regression II: Regularization

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

- ▶ Inductive biases in linear regression
- ▶ Regularization
- ▶ Model averaging
- ▶ Bayesian interpretation of regularization

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Inductive bias

- ▶ In linear regression, possible for least square solution to be non-unique, in which case there are infinitely-many solutions.
- ▶ Which one should we pick?
- ▶ Possible answer: Pick shortest solution, i.e., of minimum $\|w\|_2$

<https://eduassistpro.github.io>

change in output

(easy consequence of Cauchy-Schwarz)

▶ Note: data does not give reason to choose shorter w .

- ▶ Preference for short w is an example of an inductive bias.
- ▶ All learning algorithms encode some form of inductive bias.

Example of minimum norm inductive bias

- ▶ Trigonometric feature expansion

$$\varphi(x) = (\sin(x), \cos(x), \dots, \sin(32x), \cos(32x)) \in \mathbb{R}^{64}$$

- ▶ $n = 32$ training examples

- ▶ Infinitely many solutions to normal equations

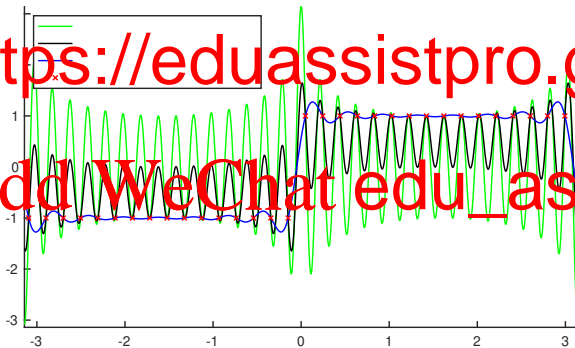


Figure 1: Fitted linear models with trigonometric feature expansion

Representation of minimum norm solution (1)

- **Claim:** The minimum (Euclidean) norm solution to normal equations lives in span of the x_i 's (i.e., in $\text{range}(A^T)$).

Assignment Project Exam Help

$$w = A^T \alpha = \sum \alpha_i x_i$$

<https://eduassistpro.github.io>

- **Proof:** If we have any solution of the form $w = s + r$, where $s \in \text{range}(A^T)$ and $r \perp \text{range}(A^T)$, and $r \neq 0$ is in $\text{null}(A)$, we can remove r and have a shorter solution:

$$A^T b = A^T A w = A^T A (s + r) = A^T A s + A^T (A r) = A^T A s.$$

(Recall Pythagorean theorem: $\|w\|_2^2 = \|s\|_2^2 + \|r\|_2^2$)

Representation of minimum norm solution (2)

- ▶ In fact, minimum Euclidean norm solution is unique!
- ▶ If two distinct solutions w and w' have the same length, then averaging them gives another solution $\frac{1}{2}(w + w')$ of shorter length.

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

- ▶ Combine two concerns: making both $\hat{\mathcal{R}}(w)$ and $\|w\|_2^2$ small
 - ▶ Pick $\lambda \geq 0$, and minimize

Assignment Project Exam Help

$$\mathcal{R}(w) + \lambda \|w\|_2^2$$

- ▶ <https://eduassistpro.github.io>
 - ▶ $\lambda = 0$ is OLS/ERM.
 - ▶ λ controls how much to pay attention to relative to data fitting term $\mathcal{R}(w)$
 - ▶ λ is hyperparameter to tune (e.g., using

Add WeChat edu_assist_pro

- ▶ Solution is also in span of the x_i 's (i.e., in $\text{range}(A^T)$)

Example of regularization with squared norm penalty

- ▶ Trigonometric feature expansion

Assignment Project Exam Help

- ▶ Trade-off between fit to data and regularizer

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

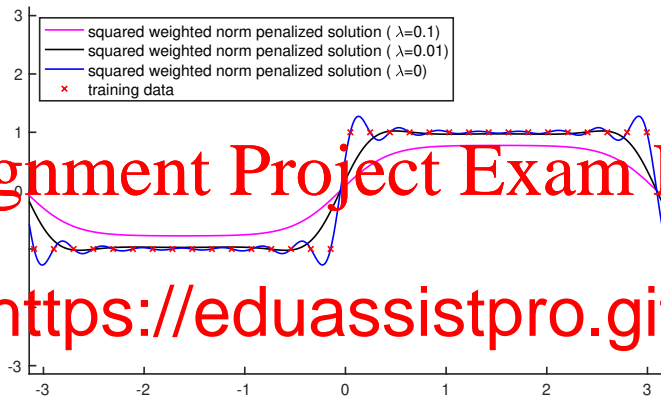


Figure 2: Trading off between data fitting and model complexity.

Data augmentation (1)

- ▶ Let $\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda}I \end{bmatrix} \in \mathbb{R}^{(n+d) \times d}$ and $\tilde{b} = \begin{bmatrix} b \\ 0 \end{bmatrix} \in \mathbb{R}^{n+d}$

▶ Then $\|\tilde{A}u - \tilde{b}\|_2^2 = \mathcal{L}(u) + \lambda \|u\|_2^2$ (ridge regression objective)

<https://eduassistpro.github.io>



- ▶ d “fake” data points, ensures augmen

- ▶ All corresponding labels are zero

- ▶ $\tilde{A}^\top \tilde{A} = A^\top A + \lambda I$ and $\tilde{A}^\top \tilde{b} = A^\top b$

- ▶ So ridge regression solution is $\hat{w} = (A^\top A + \lambda I)^{-1} A^\top b$

- ▶ Domain-specific data augmentation: e.g., image transformations

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Figure 3: What data augmentations make sense for OCR digit recognition?

- ▶ Lasso: minimize $\hat{\mathcal{R}}(w) + \lambda \|w\|_1$
 - ▶ Here, $\|v\|_1 = \sum_{j=1}^n |v_j|$, sum of absolute values of vector entries
 - ▶ Prefers short w , where length is measured using different norm
 - ▶ Tends to produce w that are sparse (i.e., have few non-zero

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Lasso vs ridge regression

- ▶ Example: coefficient profile of Lasso vs ridge
- ▶ x = clinical measurements, y = level of prostate cancer antigen
- ▶ Horizontal axis: varying λ (large λ to left, small λ to right)
- ▶ Vertical axis: coefficient value in Lasso and ridge solutions, for eight different features

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Inductive bias from minimum ℓ_1 norm

- **Theorem:** Pick any $w \in \mathbb{R}^d$ and any $\varepsilon \in (0, 1)$. Form $\tilde{w} \in \mathbb{R}^d$ by including the $\lceil 1/\varepsilon^2 \rceil$ largest (by magnitude) coefficients of w and setting remaining entries to zero. Then

$$\|\tilde{w} - w\|_2 \leq \varepsilon \|w\|_2.$$

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

- ▶ Lasso also tries to make coefficients small. What if we only care about sparsity?

- ▶ Subset selection minimize empirical risk among all k -sparse solutions



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

included variables.

- ▶ Often works as well as Lasso

- ▶ Why do we care about sparsity?

Detour: Model averaging

- ▶ Suppose we have M real-valued predictors, $\hat{f}_1, \dots, \hat{f}_M$
- ▶ How to take advantage of all of them?
- ▶ Model selection: pick the best one, e.g., using hold-out method
- ▶ Model averaging: form “ensemble” predictor \hat{f}_{avg} , where for

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu_assist_pr

Risk of model averaging

- $\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]$ for some random variable (X, Y) taking values in $\mathcal{X} \times \mathbb{R}$.

- **Theorem.** For any $f_1, \dots, f_M: \mathcal{X} \rightarrow \mathbb{R}$ the ensemble predictor $\hat{f}_{\text{avg}} := \frac{1}{M} \sum_{j=1}^M \hat{f}_j$ satisfies

<https://eduassistpro.github.io>

- Better than model selection when:

- all \hat{f}_j have similar risks, and

- all \hat{f}_j predict very differently from eac

Stacking and features

- ▶ In model averaging, “weights” of $1/M$ for all \hat{f}_j seems arbitrary
- ▶ Can “learn” weights using linear regression!
 - ▶ Use feature expansion $\phi(x) = (\hat{f}_1(x), \dots, \hat{f}_M(x))$
 - ▶ Called stacking

Assignment Project Exam Help

<https://eduassistpro.github.io>

- ▶ Upshot: Any function (even learned function)
- ▶ Conversely: Behind every feature is a deliberate choice

Detour: Bayesian statistics

- ▶ Bayesian inference: probabilistic approach to updating beliefs
 - ▶ Posit a (parametric) statistical model for data (likelihood)
 - ▶ Start with some beliefs about the parameters of model (prior)
 - ▶ Update beliefs after seeing data (posterior)

<https://eduassistpro.github.io>

- ▶ (Finding normalization constant
computationally challenging part of b
- ▶ Basis for reasoning in humans (maybe?)

Beyond Bayesian inference

- ▶ Can use Bayesian inference framework for designing estimation/learning algorithms (even if you aren't a Bayesian!)
 - ▶ E.g.: Instead of computing entire posterior distribution, find the w with highest posterior probability
 - ▶ Called maximum a posteriori (MAP) estimator

<https://eduassistpro.github.io>

- ▶ (Avoids issue with finding normalization)

Add WeChat edu_assist_pro

Bayesian approach to linear regression

- ▶ In linear regression model, express prior belief about $w = (w_1, \dots, w_d)$ using a probability distribution with density function

- ▶ Simple choice: $\text{prior}(w_1, \dots, w_d) = \prod_{j=1}^d \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{w_j^2}{2\sigma^2}\right)$ variables

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

Add WeChat edu_assist_pr

MAP for Bayesian linear regression

- Find w to maximize

$$\prod_{j=1}^d \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{w_j^2}{2\sigma_j^2}\right) \prod_{i=1}^n p(x_i) \frac{1}{\sqrt{1\pi}} \exp\left(-\frac{(y_i - x_i^\top w)^2}{1}\right).$$

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

Add WeChat `edu_assist_pro`

- For $\sigma^2 = \frac{1}{n\lambda}$, same as minimizing

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

which is the ridge regression objective!

Example: Dartmouth data example

- ▶ Dartmouth data example, where we considered intervals for the HS GPA variable:

$(0.00, 0.25]$, $(0.25, 0.50]$, $(0.50, 0.75]$, \dots

- ▶ <https://eduassistpro.github.io>
- ▶ College GPA values.
- ▶ What's the Bayesian interpretation of mi objective?

$$\frac{1}{n} \sum_{i=1}^n (\varphi(x_i)^\top w - y_i)^2 + \lambda \sum_{j=1}^d (w_j - \mu)^2$$