CMSC5741 Big Data Tech. & Apps.

Lecture 9: Large Scale

Su ines

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Prof. Michael R. Lyu

Computer Science & Engineering Dept.

The Chinese University of Hong Kong

Motivation

- Introduce the widely used classification tool: Support Vector Machine (SVM)

Assignment Project Exam Help

- Understand t eter estimation method in ter

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

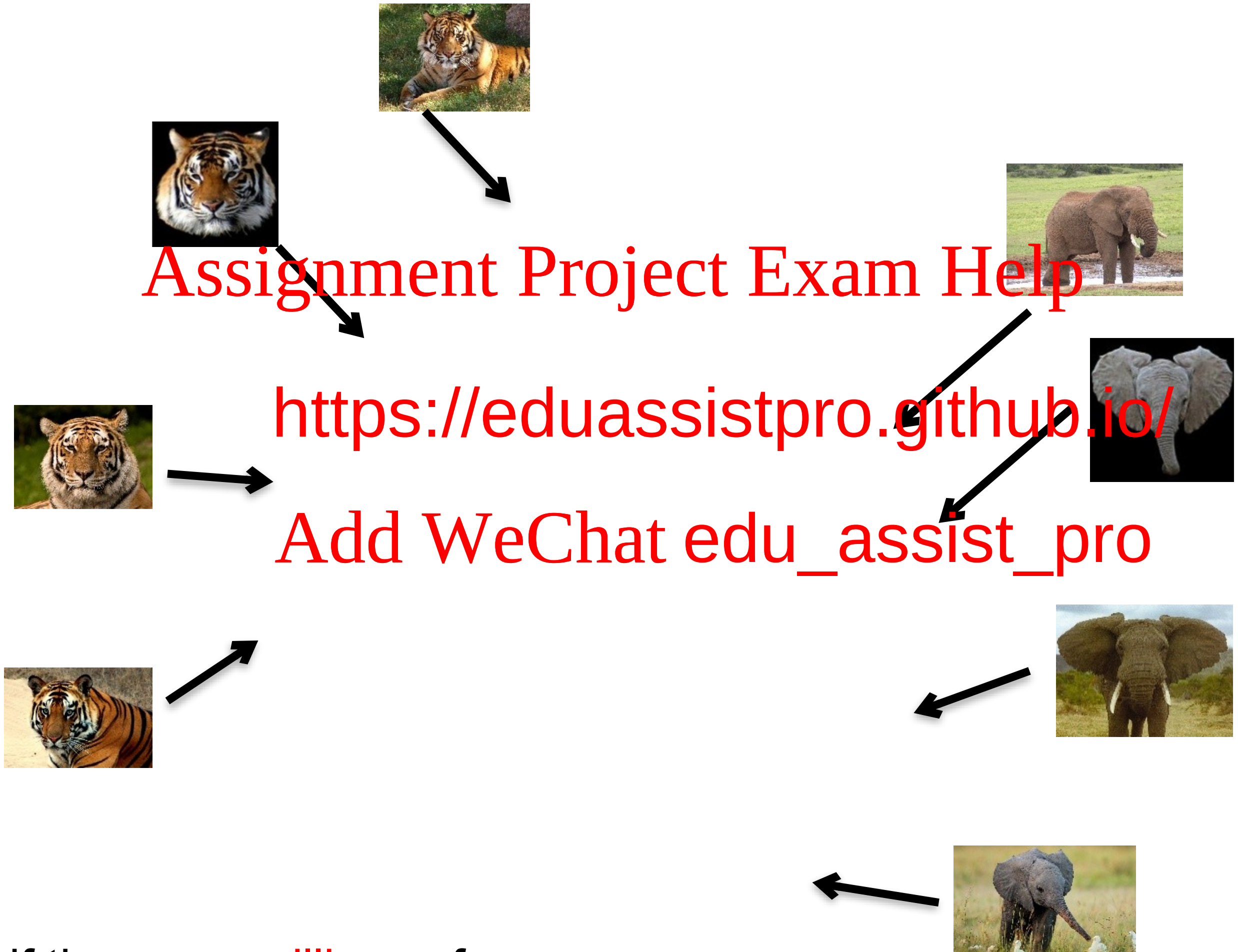
Motivation

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Motivation



Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

What if there are millions of photos, how to make the SVM training scalable?

Outline

- Support Vector Machines
  - History
  - Linear Separa
  - Non-linear Se
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Outline

- Support Vector Machines
  - History
  - Linear Separa
  - Non-linear Se
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

SVMs: History

- SVMs introduced in COLT-92 by Boser, Guyon & Vapnik. Became rather popular since.

Assignment Project Exam Help

- Theoretically rithm: developed from Statistic

https://eduassistpro.github.io/

apnik & Chervonenkis) since the 6

Add WeChat edu_assist_pro

- Empirically good performance: successful applications in many fields (bioinformatics, text, image recognition, . . .)

SVMs: History

- Centralized website: www.kernel-machines.org.
- Several textbooks, e.g. "An introduction to Support Vecto tianini and Shawe-Taylor
- A large and diverse comm rk on them: from machine learning, optimization, statistics, neural networks, functional analysis, etc.

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Outline

- Support Vector Machines
  - History
  - Linear SVMs
  - Non-linear SVMs
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Linear SVMs

- Data
  - Training examples: $(x_1, y_1), \ldots, (x_n, y_n)$

Assignment Project Exam Help

  - Each
  - Want to fi

https://eduassistpro.github.io/

    to separate "+" from "-"

Add WeChat edu_assist_pro

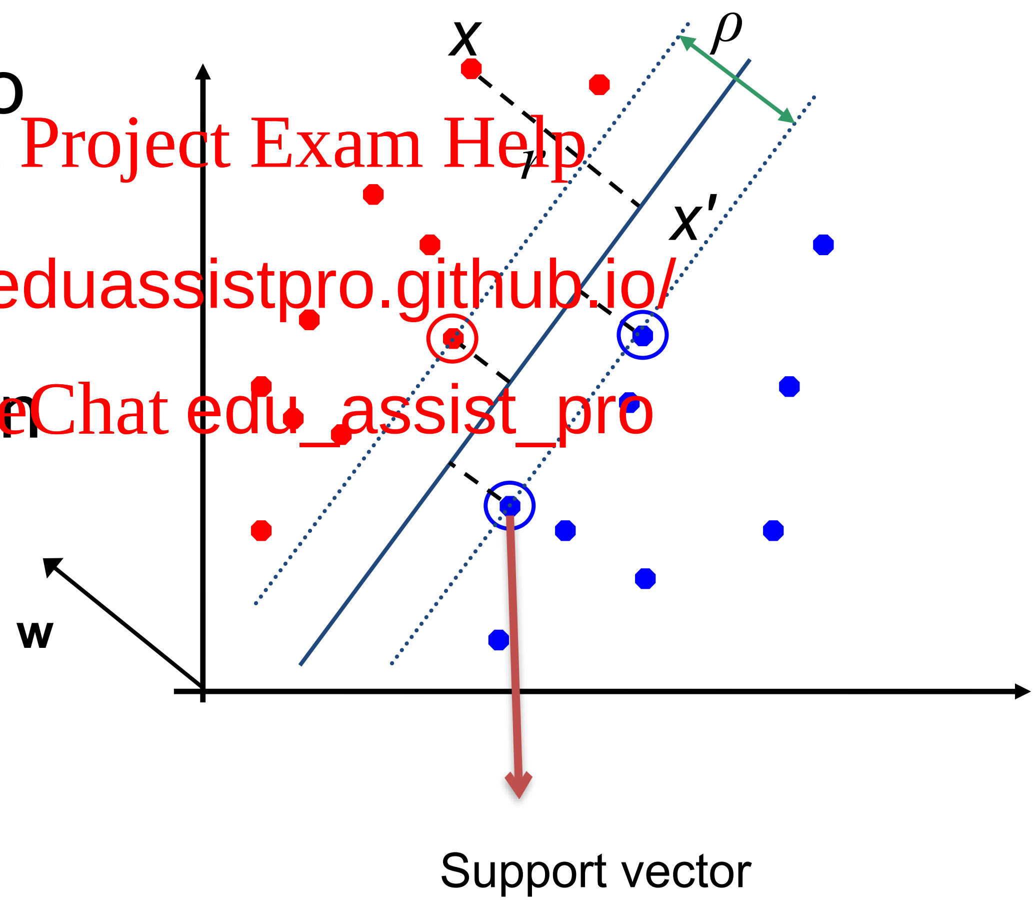
- What's the best hyperplane defined by w ?

Largest Margin

- Distance from the separating hyperplance corresponds to the “confide
Assignment Project Exam Help
  prediction
https://eduassistpro.github.io/
- Example: We have more confidence to say A and B belong to “+” than C
Add WeChat edu_assist_pro

Largest Margin

- **Support Vectors**: Examples closest to the hyperpl
- **Margin** ρ : separation between support vectors of classes.



Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Largest Margin

- Distance from example to the separator is :

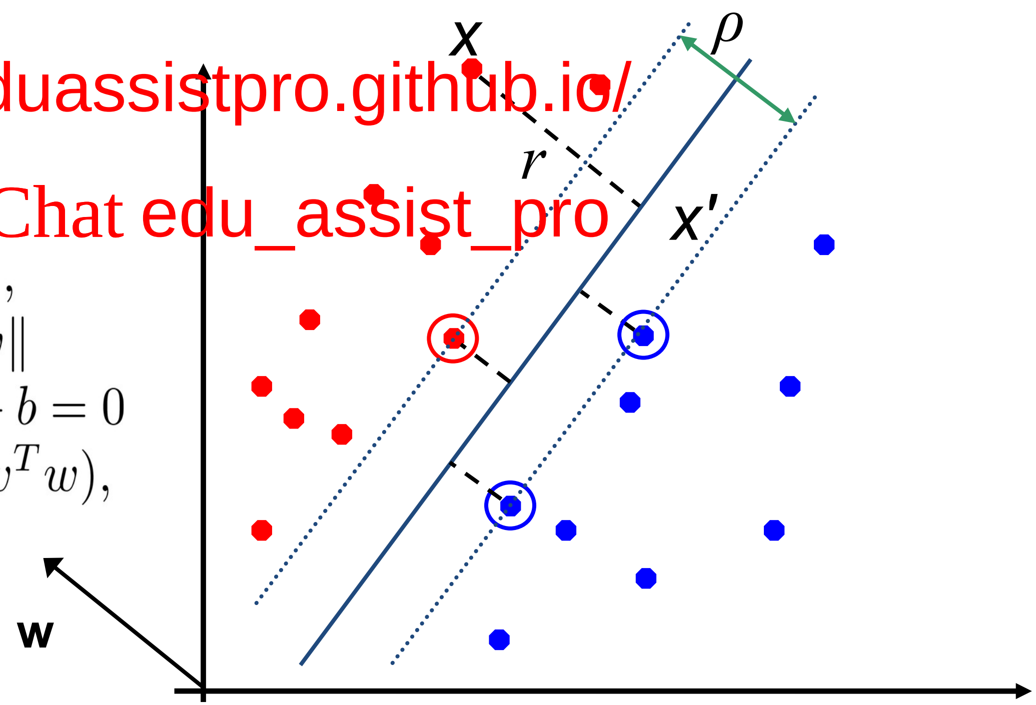
Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

- Proof:

$x' - x // w$, unit vector is $w/\|w\|$,
so line is $rw/\|w\|$, $x' = x - yrw/\|w\|$
since x' is on the separator, $w^T x' + b = 0$
so $w^T(x - yrw/\|w\|)$+b=0,||w||= $\sqrt{(w^T w)}$,
so $w^T x - yr\|w\| + b = 0$,
then we get $r = y\frac{w^T x + b}{\|w\|}$



Largest Margin

- Assume that all data is at least distance 1 from the hyperplane, then the following constraints follow for a training

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

- For support vectors, the inequality becomes an equality
- Recall that $$r = y\frac{w^T x + b}{\|w\|}$$
- Margin is: $$\rho = \frac{2}{\|w\|}$$

Linear SVMs

- Note that we assume that all data points are linearly separated by the hyperplane.

Assignment Project Exam Help

- The margin is of parameters.
  - i.e. by changi margin doesn't change

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Linear SVMs

- Maximize the margin
  - Good according to intuition, theory (VC dimension) & practice

Assignment Project Exam Help

- The problem https://eduassistpro.github.io/ rmulated as:

Add WeChat edu_assist_pro

- An equivalent form is:

$$\min_{w} \frac{1}{2}\|w\|^2$$
$$s.t. \quad y_i(w^T x_i + b) \geq 1 \quad \forall i = 1, \ldots, n$$

Outline

- Support Vector Machines
  - History
  - Linear Separa
  - Non-linear Se
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Non-Linear Separable SVMs

- In reality, training samples are usually not linearly separable.

Assignment Project Exam Help

- Soft Margin

https://eduassistpro.github.io/

  - Idea: allow slack variable ξ_i to errors

Add WeChat edu_assist_pro

  - Still try to minimize training set errors, and to place hyperplane "far" from each class (large margin)

Soft Margin Classification

- The problem becomes:

Assignment Project Exam Help

https://eduassistpro.github.io/

  - Minimize $\|w\|^2$ aining mistakes
  - Set C using cross validation

Add WeChat edu_assist_pro

Soft Margin Classification

- If point x_i is on the wrong side of the margin then get pe

Assignment Project Exam Help

- Thus all mis

https://eduassistpro.github.io/

equally bad!

Add WeChat edu_assist_pro

Slack Penalty C

$$\min_{w} \frac{1}{2}\|w\|^2 + C\sum \xi_i$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i = 1, \ldots, n$$

Assignment Project Exam Help

- What is the C:

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

  - $C = 0$: can set ξ_i to anything, then w=0 (basically ignore the data)
  - $C = \infty$: Only want w, b to separate the data

Soft Margin Classification

- SVM in the "natural" form

Assignment Project Exam Help

Marg

https://eduassistpro.github.io/

pirical **loss L**

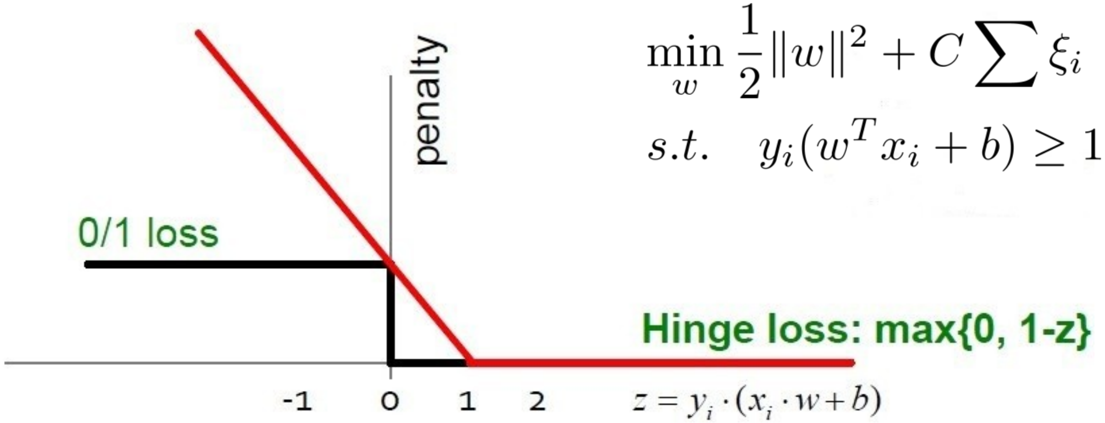
Parameter

Add WeChat edu_assist_pro

- SVM uses "Hinge Loss":

$$\min_{w} \frac{1}{2}\|w\|^2 + C\sum \xi_i$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i = 1, \ldots, n$$



In-class Practice

- Go to practice

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Outline

- Support Vector Machines
  - History
  - Linear SVMs
  - Non-linear SVMs
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

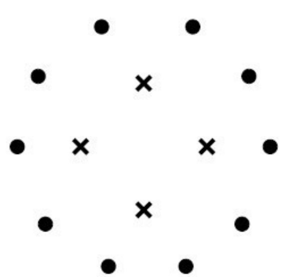
Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Non-linear Separable SVMs

- Linear classifiers aren't complex enough sometimes.
Assignment Project Exam Help
  - Map data int e including non-linear feature
https://eduassistpro.github.io/
  - Then construct a hyperpla pace so all other equations are the same
Add WeChat edu_assist_pro



Non-linear Separable SVMs

- Formally, process the data with:

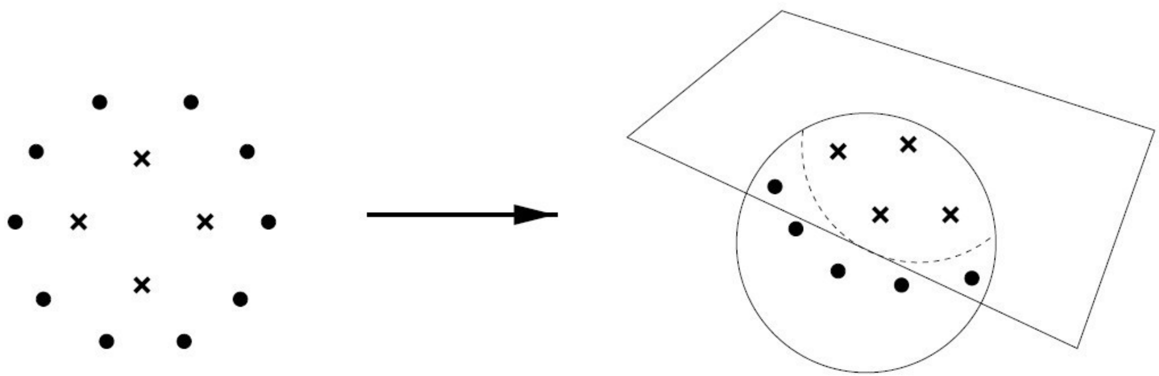
$$x \mapsto \Phi(x)$$

Assignment Project Exam Help

- Then learn th y

https://eduassistpro.github.io/

Add WeChat edu_assist_pro



Example: Polynomial Mapping

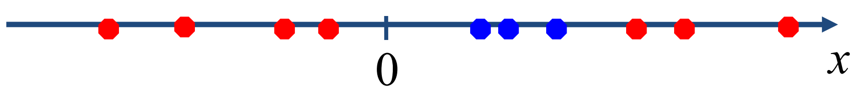
$$\Phi : R \rightarrow R^2$$

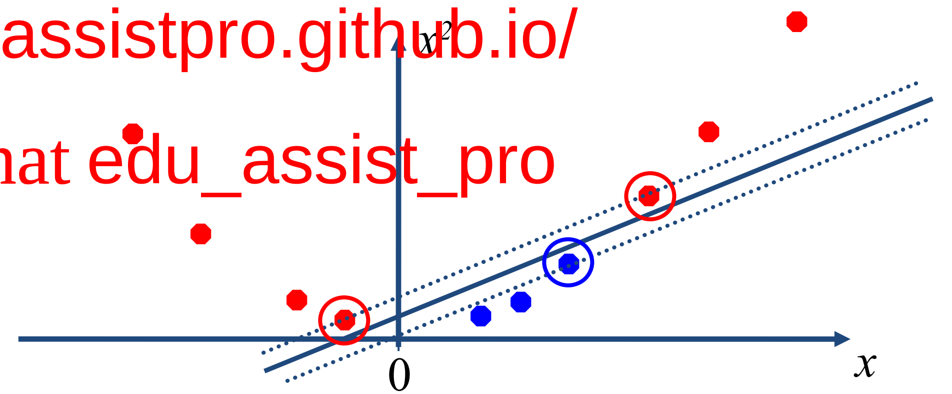
$$(x_1) \mapsto (z_1, z_2) := (x_1, x_1^2)$$

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro





Example: Polynomial Mapping

$$\Phi : R^2 \to R^3$$

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$$

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Example: MNIST

- Data: 60,000 training examples, 10000 test examples, 28x28
Assignment Project Exam Help
- Linear SVM ha error. Polynomial SVM has arou
https://eduassistpro.github.io/

Add WeChat edu_assist_pro

MINST Results

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Choosing a good mapping $\Phi(\cdot)$ (encoding prior knowledge + getting right complexity of function class) for your problem improves results.

SVMs: Kernel Trick

- The Representer theorem (Kimeldorf & Wahba, 1971) shows that (for SVMs as a special case):

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

  for some variables α, inst timizing w directly, we can optimize α.

- The decision rule is: $f(x) = \sum_{i=1}^{m} \alpha_i \Phi(x_i) \cdot \Phi(x) + b$
  - We call $K(x_i, x) = \Phi(x_i) \cdot \Phi(x)$ the *kernel function*.

Kernels

- Why kernels?
  - Make non-separable problem separable.
  - Map data int nal space

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

- Common used
  - Linear
  - Polynomial $K(x_i, x_j) = (1 + x_i^T \cdot x_j)^d$
    - Gives feature conjunctions
  - Radial basis function

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2}$$

Outline

- Support Vector Machines
  - History
  - Linear Separa
  - Non-linear Se
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

SVM: How to Estimate w, b

- We take the soft margin classification for example:

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

- Standard way
  - Solver: software for finding to "common" optimization problems, e.g. LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)
- Problems: Solvers are inefficient for big data!

SVM: How to Estimate w, b

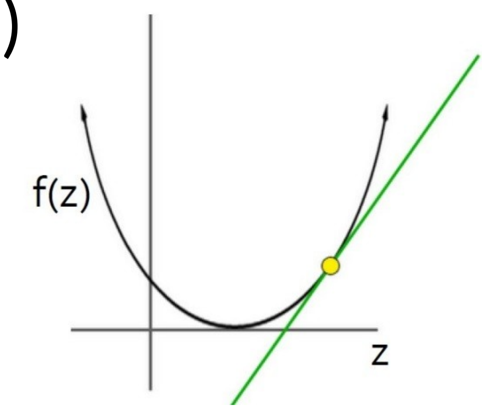
- Want to estimate w, b !

$$\min_w \frac{1}{2}\|w\|^2 + C\sum \xi_i$$

$$s.t. \; \forall i \; y_i(w^T x_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0$$

- Alternative approach:
  Assignment Project Exam Help
  – Want to mini
  https://eduassistpro.github.io/

  Add WeChat edu_assist_pro
  – How to minimize convex functions f(z)
  – Use gradient descent: $\min_z f(z)$
  – Iterate: $z_{t+1} \leftarrow z_t - \eta f'(z_t)$



SVM: How to Estimate w?

- Want to minimize ***f(w,b)***:

Assignment Project Exam Help

https://eduassistpro.github.io/

irical **loss L**

Add WeChat edu_assist_pro

- Compute the gradient

$$\nabla(j) = \frac{\partial f(w,b)}{\partial w^{(j)}} = w^{(j)} + C\sum_{i=1}^{n}\frac{\partial L(x_i, y_j)}{\partial w^{(j)}}$$

$$\frac{\partial L(x_i, y_j)}{\partial w^{(j)}} = \begin{cases} 0 & \text{if } y_i(w \cdot x_i + b) \geq 1 \\ -y_i x_i^{(j)} & \text{otherwise} \end{cases}$$

SVM: How to Estimate w?

- Gradient descent:

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

- Problem:
  - Computing $\nabla_{(j)}$ takes O(n) time
    - n … size of the training dataset

SVM: How to Estimate w?

We just had:
$\nabla(j) = w^{(j)} + C \sum_{i=1}^{n} \frac{\partial L(x_i, y_i)}{\partial w^{(j)}}$

- Stochastic Gradient Descent
  - Instead of evaluating gradient over all examples, evaluate it for each individual training example

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

- Stochastic gradient desce

Iterate untial convergence:

- For $i = 1, \ldots, n$
  - For $j = 1, \ldots, d$
    - Evaluate: $\nabla(j, i)$
    - Upadate: $w^{(j)} \leftarrow w^{(j)} - \eta \nabla(j, i)$

Example: Text Categorization

- Example by Leon Bottou:
  - **Reuters RCV1** document corpus

Assignment Project Exam Help

    - Predict a category of a document
      - One **vs.** th

https://eduassistpro.github.io/

  - ***n = 781,000*** training exam ents)

Add WeChat edu_assist_pro

  - 23,000 test examples
  - **d = 50, 000** features
    - One feature per word
    - Remove stop-words
    - Remove low frequency words

Examples: Text Categorization

- Questions:
  - Is SGD successful at minimizing *f(w,b)*?
  - How quickly of *f(w,b)*?
  - What is the e

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

- SGD-SVM is successful at minimizing the value of *f(w,b)*
- SGD-SVM is super fast
- SGD-SVM test set error is comparable

Optimization “Accuracy”

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

SGD vs. Batch Conjugate Gradient

- SGD on full dataset vs. Batch Conjugate
  - Gradient on a sample of ***n*** training examples

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Practical Considerations

- Need to choose learning rate η and t_0

$$w_{t+1} \leftarrow w_t - \frac{\eta_t}{t + t_0}\left(w_t + C\frac{\partial L(x_i, y_i)}{\partial w}\right)$$

Assignment Project Exam Help

- Leon suggests

https://eduassistpro.github.io/

  – Choose t_0 so al updates are comparable with the expe f the weights

Add WeChat edu_assist_pro

  – Choose η:
    - Select a small subsample
    - Try various rates η (e.g., 10,1,0.1,0.01,…)
    - Pick the one that most reduces the cost
    - Use η for next 100k iterations on the full dataset

Practical Considerations

- Sparse Linear SVM:
  - Feature vector x_i is sparse (contains many zeros)
    - Do not do:
    - But represe $\mathbf{x_i} = [(\mathbf{4}, \mathbf{1}), (\mathbf{9}, \mathbf{5}), \ldots]$
  - Can we do the SGD update iently?

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

  - Approximated in 2 steps:

$$w \leftarrow w - \eta C \frac{\partial L(x_i, y_i)}{\partial w}$$

$$w \leftarrow w(1 - \eta)$$

Cheap: $\mathbf{x_i}$ is sparse and so few coordinates **j** of **w** will be updates
Expensive: **w** is not sparse, all coordinates need to be updated

Practical Considerations

- Solution 1: $\mathbf{w} = \mathbf{s} \cdot \mathbf{v}$
  - Represent vector **w** as the product of scalar **s** and the vector **v**

Assignment Project Exam Help

  - Then the u

https://eduassistpro.github.io/

    - 1) $v = v - \eta C \frac{\partial L(x_i, y_i)}{\partial v}$

Add WeChat edu_assist_pro

    - 2) $s = s(1 - \eta)$
- Solution 2:
  - Perform only step 1) for each training example
  - Perform step 2) with lower frequency and higher η

Practical Considerations

- Stopping criteria:

How many iterations of SGD?

Assignment Project Exam Help

– Early st dation

https://eduassistpro.github.io/

  - Create valid
  - Monitor cost function on th set

Add WeChat edu_assist_pro

  - Stop when loss stops decreasing

Practical Considerations

- Stopping criteria:

  How many iterations of SGD?

  – Early Stoppin

    - Extract two **B** of training data
    - Train on **A**, stop by validatin
    - Number of epochs is an estimate of **k**
    - Train for **k** epochs on the full dataset

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

What about Multiple Classes?

- Idea 1:
  - One against all

Assignment Project Exam Help

Learn 3 classifi

- + vs. {o,-}
- - vs. {o,+}
- o vs. {+,-}

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Obtain: $w_+ b_+, w_- b_-, w_o b_o$

– Return class *c*

$$\arg\max_c w_c x + b_c$$

What about Multiple Classes?

- Idea 2:
  - Learn 3 sets of weights simultaneously
  - Want the cor est margin:

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Multiclass SVM

- Optimization problem:

Assignment Project Exam Help

https://eduassistpro.github.io/

  – To obtain par class c, we can use similar techniques as for 2

Add WeChat edu_assist_pro

- SVM is widely perceived a very powerful learning algorithm

Demo

Libsvm package for R:

Assignment Project Exam Help

http://cran.r-project.org/web/packages/e1071/index.html

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Demo

```
> # load library, class, a dependence for the SVM library
> library(class)
> # load library, SVM
> library(e1071)
> # load library, mlbench, a collection of some datasets from the UCI repository
> library(mlbench)
> # load data, has 7 classes,
    http://archive.ics.uci.edu/ml/datasets/Glass+Id
> data(Glass, package = "mlbench")
> # get the index of all data
> index <- 1:nrow(Glass)
> # generate test index
> testindex <- sample(index, trunc(length(index)/3))
> # generate test set
> testset <- Glass[testindex, ]
> # generate trainin set
> trainset <- Glass[-testindex, ]
```

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Demo

```
> # train svm on the training set
> # cost=100: the penalizing parameter for C-classification
> # gamma=1: the radial basis function-specific kernel parameter
> # Output values include SV, index, coefs, rho, sigma, probA, probB
> svm.model <- svm(Type~ ., data = trainset, cost = 100, gamma = 1)
> # a vector of predicted val
> # for classification: a vecto
> svm.pred <- predict(svm.model, testset[, -10])
> # a cross-tabulation of the true
> # versus the predicted values
> table(pred = svm.pred, true = testset[, 10])
```

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

One-slide Takeaway

- SVM:
  - Linear Separable SVMs
  - Non-linear Se argin and Kernel Trick

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

- Parameter Estimation:
  - Solver: e.g. libsvm, not efficient
  - Stochastic gradient descent

Outline

- Support Vector Machines
  - History
  - Linear Separa
  - Non-linear Se
    - Soft Margin
    - Kernel Trick
- Parameter Estimation
- Further Reading

Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro

Further Reading

- Early paper about SVM algorithm: http://link.springer.com/content/pdf/10.1007%2FBF00994018.pdf Assignment Project Exam Help
- More kernel t https://eduassistpro.github.io/
  - Schölkopf, Bernhard; Burg Add WeChat edu_assist_pro pher J. C.; and Smola, Alexander J. (editors); *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1999. ISBN 0-262-19416-3.

Further Reading

- More efficient learning algorithm for SVM:
  - Parallelizing Support Vector Machines on Distributed Computers: https://code.google.com/p/psvm/

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Reference

- http://www.stanford.edu/class/cs246/slides/13-svm.pdf
- http://www.stanford.edu/class/cs276/handouts/lecture14-SVMs.ppt
- http://i.stanford.e
- http://www.svms.
- http://www.cs.columbia.edu/~kat ocuments/jason_svm_tutorial.pdf
- http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- Chang, E, Zhu, K, Wang, H, Bai, H, Li, J, Qiu, Z, and Cui, H. PSVM: Parallelizing support vector machines on distributed computers. NIPS, 20:257-264. 2007.

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

In-class Practice

(2,3)

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

- Consider building an SVM over the (very little) data set shown in above figure, compute the each SVM decision boundary.