SMAI-M20-L37: Programming and Practice of NN in PyTorch

C. V. Jawahar

IIIT Hyderabad

November 16, 2020

Announcements

1. Winding up CRs and Home works. We have buffer for specific difficulties (like power cuts).
2. Quiz 3: Academic office has agreed with 25, 9.30-10.30 (It will be Similar to Q2). Scope: Topics after Q1.
3. Absence/Requests/Special Issues from Q1 and Q2: Will be addressed after Q3. ($<$ 5 students?)

Class Review

Programming in PyTorch

Blank

Recap:

- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization (viii) Momentum (ix) Convolutional Layer (x) Recurrent/Feedback networks (xi) Auto-encoder and unsupervised learning
- **Beyond Simple Supervised Learning** (i) Paradigms of Learning (ii)

Practice of Neural Networks

1. Programming in PyTorch
2. Defining Neural Networks
3. Training Neural Networks
4. Overfitting
    1. Data Augmentation
    2. L1 and L2 Regularization
    3. Dropout
5. Fine Tuning
    1. Re-Use of the intermediate representations $\phi(x)$.
    2. Fine tuning/Refining for a new task.

This Lecture: In Class Problem Solving

1. Discuss within the class, Ask queries and clarifications.
2. Solve a new problem (where almost all the code is available from the previous notebook)
    1. `https://colab.research.google.com/drive/1EtlSmUAcHc8eLtaOMtvCuG_ppMkDxH8a?usp=sharing`
    2. Submit the link to the notebook, pdf version of the notebook
    3. Write a brief report of what you observe and submit. (similar to the problem we solved on paper). (submit pdf; one or two pages; If appropriate, keep graphs and conceptual explanations).

SMAI-M20-L36: Programming and Practice of NN in PyTorch

C. V. Jawahar

IIIT Hyderabad

November 13, 2020

Class Review

Programming in PyTorch

Blank

Recap:

- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization (viii) Momentum (ix) Convolutional Layer (x) Recurrent/Feedback networks (xi) Auto-encoder and unsupervised learning
- **Beyond Simple Supervised Learning** (i) Paradigms of Learning (ii)

This Lecture:

1. See the two associated videos:
   1. https://youtu.be/s6aVkulQgd0
   2. https://youtu.be/xRa-hO54dh0
2. Learn how to work with the following associated note books:
   1. run on collab ?
   2. https://colab.research.google.com/drive/1leWjPMRbIW4_4W8yVCvliYEceiVF6DPv?usp=sharing
   3. https://colab.research.google.com/drive/1OWOhsRO2LKIEHFUiLMJn1Sea-qc_tFgv?usp=sharing
3. Ask queries and clarifications.

SMAI-M20-L35: Programming for ML/MLP/NN in Recent PyTorch

C. V. Jawahar

IIIT Hyderabad

November 11, 2020

Class Review

1. K-Means:
   1. What K-Means does, guarantees?
   2. Running K-Means on a numerical example
   3. What influences K-Means performance?

Blank

Recap:

- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization (viii) Momentum (ix) Convolutional Layer (x) Recurrent/Feedback networks (xi) Auto-encoder and unsupervised learning
- **Beyond Simple Supervised Learning** (i) Paradigms of Learning (ii)

This Lecture:

1. See the two associated videos:
    1. https://youtu.be/u3rUkkh-Rac
    2. https://youtu.be/xnJuNxURtik
2. Learn how to work with the following associated note books:
    1. run on collab ?
    2. https://colab.research.google.com/drive/1SfldyIXsE-oUds_GC2z9pioE4Ovezhzg?usp=sharing
    3. https://colab.research.google.com/drive/1xIBq2bMNaCb7LxLq7TYyhGKalfgye3hX?usp=sharing
3. Ask queries and clarifications.

SMAI-M20-L34: Beyond Supervised Learning

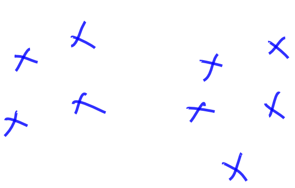
C. V. Jawahar

IIIT Hyderabad

November 9, 2020

Class Review

"Spectral clustering"



1. Consider a hierarchical classifier
   - Create an MST
   - Successively remove the longest or largest edges.

   What objective does it optimizes? maximizes? minimises? Is the answer unique? Does the performance depend on initialization?
2. Is clustering unique? Where do we use clustering?

$$\begin{bmatrix} d(x_i x_j) & & \\ & & \\ & & \end{bmatrix}_{N \times N}$$

N - vertices

all pairs

Blank

Recap:

- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization (viii) Momentum (ix) Convolutional Layer (x) Recurrent/Feedback networks (xi) Auto-encoder and unsupervised learning
- **Beyond Simple Supervised Learning** (i) Paradigms of Learning (ii)

This Lecture:

1. **K-Means Clustering**
   1. Initialize Clusters randomly
   2. Compute Means for each cluster
   3. Re-Assign samples based on the distances to the means
   4. Repeat until no-change
2. **Self-Training**
   1. Small number of Labelled and Large number of unlabelled examples
   2. Build a classififer with labelled examples.
   3. Predict on unlabelled and use these “pseudo-labels” as “labells” and train the new classifier.
   4. Take only selected (say most confident samples for re-training).

Questions? Comments?

Blank

Blank

Blank

Discussions Point - I

Consider K-Means (with K=2)

1. Six 1D samples
$$\{-3, -2, -1, +1, +2, +3\}$$
with $C_1 = \{-3, -2, +1\}$ and $C_2 = \{+3, +2, -1\}$
2. Six 1D samples
$$\{-3, -2, -1, +1, +2, +3\}$$
with $C_1 = \{-3, +2, +1\}$ and $C_2 = \{+3, -2, -1\}$
3. Four 2D samples:
$$(-1, -1), (-1, +1), (+1, +1), (+1, -1)$$
with initialization as:
   - $C_1 = \{(-1, -1), (-1, +1)\}$ and $C_2 = \{(+1, -1), (+1, +1)\}$
   - $C_1 = \{(+1, +1), (-1, +1)\}$ and $C_2 = \{(+1, -1), (-1, -1)\}$
4. Comment on the optimization problem that K-Means solves (i) sensitivity to initialization (ii) convergence (iii) Local and global minima etc.

Blank

Blank

Blank

Blank

Blank

Discussions Point -II

Evaluation of Clustering is More Challenging than Classification. Objective is to sensitize the practical issues than exploring fully here.

How do we do performance evaluation in the Clustering problems?

1. How do we evaluate/quantify the quality of a specific cluster?
2. Does the objective (value) of the K-Means Obj. fn act as a good measure? How does it change with K? What should be desirable?
3. How do we evaluate/quantify the computational complexity of clustering?
4. How do we evaluate the final clusters (results)
    1. When there is no ground truth (GT)?
    2. Compare two different outputs, and choose the better one?
    3. When K is same (in both) and different?
    4. When there is full GT (as if from a manually labelled data as in classification; may be only in some course assignments!!)
    5. When there is some partial GT (some samples were manually labelled)
    6. When there is some partial GT (some pairs were labelled as "similar" and some "dissimilar". (or "must link", and "can not link" in graph terminologies)

Suggest situations where some of these will be needed.

Blank

Blank

Blank

Blank

What Next:?

1. Programming for Deep Learning.
2. NN Architectures and NN Learning (winding up)
3. Beyond Simple Supervised Learning

SMAI-M20-L33: Intro. to Feedback Networks; and Clustering

C. V. Jawahar

IIIT Hyderabad

November 6, 2020

Class Review

1. What is the a good initialization? (or what is a bad initialization[1])
2. What happens during training for an MLP?

[1] Read, Tryout and Appreciate: https://www.deeplearning.ai/ai-notes/initialization/

Blank

Recap:

- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization (viii) Momentum (ix) Convolutional Layer (x) Auto-encoder and unsupervised learning
- **Beyond Simple Supervised Learning** (i) Paradigms of Learning

This Lecture:

1. **Feedback/Recurrent Networks**
    1. Feedforward vs Feedback network
    2. Data as a sequence (of vectors)
    3. Recurrent Model
2. **Problem of Clustering**
    1. Notion of a Cluster (vs Class in supervised Learning)
    2. Hirarchical Approaches
    3. Agglomerative and Divisive (bottom-up and top-down)
3. Reading Material for Neural Networks [2]

Questions? Comments?

[2] Read in detail: https://www.dropbox.com/s/g9vu0dollo6sr48/nn-notes.pdf?dl=0

Blank

Blank

Blank

Discussions Point - I

Consider a recurrent/feedback model of neural network given by

$$s_t = f(Ux_t + Ws_{t-1})$$

$$o_t = g(Vs_t)$$

1. Why do one say that such networks have "infite or long term memory"? (or never forgets) Or RNNs capture long term dependencies?
2. Why the problem of vanishing gradient is serious in RNNs?

Blank

Blank

Blank

Discussions Point -II

Consider two Clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. As part of merging/comparing clusters (say while designing a bottom-up or agglomerative clustering) we want to compare or find similarity between clusters. What do you think of the following functions[3]?

1. $\min_{x \in \mathcal{C}_1, x_2 \in \mathcal{C}_2} s(x_1, x_2)$
2. $\max_{x \in \mathcal{C}_1, x_2 \in \mathcal{C}_2} s(x_1, x_2)$
3. $Average_{x \in \mathcal{C}_1, x_2 \in \mathcal{C}_2} s(x_1, x_2)$

Any alternate suggestions?

[3] https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Blank

Blank

Blank

Blank

What Next:?

1. NN Architectures and NN Learning (winding up)
2. Programming for Deep Learning.
3. Beyond Simple Supervised Learning

SMAI-M20-L32: Unsupervised Learning in Neural Networks

C. V. Jawahar

IIIT Hyderabad

November 2, 2020

Announcement

1. Quiz in the regular class slot on Wed. Similar to Q1/CR.
2. Topics: Topics that are not covered in Q1.
3. NN Learning (included): MLP and Back Propagation
4. NN Learning (not included): Convolution layer, Auto encoder, Momentum

Class Review

1. Convolution layer in 1D CNNs:
    1. *M* channels in the input and *N* channels in the output.
    2. Number of learnable weights or parameters
    3. Stride and impact on output size
    4. Zero padding and impact on output size.

Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization (viii) Momentum (ix) Convolutional Layer

This Lecture:

1. **Auto Encoders**
    1. Reconstruct itself with a constrained (“bottleneck”) architecture.
    2. Role as:
        1. Data Compression
        2. Unsupervised Feature Learning
        3. Non-Linear Dimensionality Reduction
    3. Comment about Popular Encoder-Decoder architectures of today.
2. **Beyond Supervised Learning**
    1. Supervised Vs Unsupervised Learning
    2. Unsupervised as “Clustering”, ”Discovery of the structure”
    3. Semi-Supervised Learning
    4. Self-Supervised Learning

Questions? Comments?

Blank

Blank

Blank

Discussions Point - I

Consider an auto enoder with fully connected layer and the architecture as:

$$1000 - 100 - 10 - 5 - 10 - 100 - 1000$$

1. Assume the activations are linear, show how this is similar to "PCA" or linear dimensionality reduction that we are familiar with.
2. If we use it for compression (say for a speech or image signal of size 1000), what can we say about the compression ratio? Why is this a good idea? Why is this a bad idea?

Blank

Blank

Blank

Discussions Point -II

We know the learning rule as:

$$w^{k+1} \leftarrow w^{k} - v_k$$

$$v_k = \eta \nabla J + \beta v_{k-1}$$

1. What if $\eta = 1.0$ and $\beta = 0.0$?
2. What if $\eta = 0.0$ and $\beta = 1.0$?
3. Assume $\nabla J = -0.1$ for all $k = 0, 1, \ldots, 10$. $v_{-1} = 0$. $w^0 = 0.0$. $\eta = 0.1$ and $\beta = 0.9$, what happens to w^k for $k = 1, 2, 3$?
4. Where it should have reached with and without momentum for $k = 10$? (Appreciate how momentum helps in speeding up, if we have a consistent slow slope).

Blank

Blank

Blank

Blank

What Next:?

1. NN Architectures and NN Learning
2. Programming for Deep Learning.
3. Beyond Simple Supervised Learning

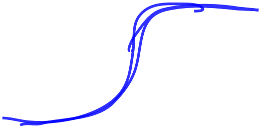
SMAI-M20-L31: Learning in Neural Networks

C. V. Jawahar

IIIT Hyderabad

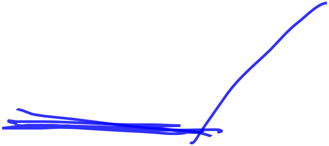
October 28, 2020

Class Review



1. ReLu activation in Neural Networks; Its properties, its derivative, its limitations, etc.

$\max(0, x)$



Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc. (i) Loss Functions and Optimization (ii) Probabilistic View, Bayesian View, MLE (iii) Eigen Vector based optimization (iv) Gradient Descent: Stochastic and Batch GD (v) Classification and Regression
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) Decision Trees (vi) SVMs (hard margin, soft margin, kernel) (vii) Kernel trick and kernelized algorithms
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP (iv) Backpropagation (v) Chain rule (vi) Activations (vii) challenges in optimization

This Lecture:

1. **Convolution Layer**
   1. A very popular layer, specially for image, speech and text data. (where local context is very important)
   2. Connection to signal processing, filtering.
   3. Do read[1]
   4. (out of scope)But strongly advised to read[2]
2. **Momentum**
   1. Improvement over gradient descent updates
   2. May read later [3]

Questions? Comments?

[1] https://danieltakeshi.github.io/2019/03/09/conv-matmul/

[2] https://cs231n.github.io/convolutional-networks/

[3] http://d2l.ai/chapter_optimization/momentum.html

Blank

Blank

Blank

Discussions Point - I

Consider an input of $1, 2, -3, 2, -1, 0, -4, 6$. Assume zero padding.

1. Assuming the layer as convolution-layer with weights as $1, 1, 1$, find the convolution output.
2. Consider two convolution weights (i) $1, 1, 1$ and (ii) $1, 0, -1$. All the neurons have ReLU activation. Find the outputs.
3. Consider a stride of two [4]. Find the outputs?

[4] When the stride is 1 (that is what we did till now), then we move the filters by 1 sample at a time. When the stride is 2 then we move the filters by 2 samples. Effectively output is half the size of the input.

Blank

Blank

Blank

Discussions Point -II

We know the learning rule as:

$$w^{k+1} \leftarrow w^{k} - v_k$$

$$v_k = \eta \nabla J + \beta v_{k-1}$$

1. What if $\eta = 1.0$ and $\beta = 0.0$?
2. What if $\eta = 0.0$ and $\beta = 1.0$?
3. Assume $\nabla J = -0.1$ for all $k = 0, 1, \ldots, 10$. $v_{-1} = 0$. $w^0 = 0.0$. $\eta = 0.1$ and $\beta = 0.9$, what happens to w^k for $k = 1, 2, 3$?
4. Where it should have reached with and without momentum for $k = 10$? (Appreciate how momentum helps in speeding up, if we have a consistent slow slope).

Blank

Blank

What Next:?

1. NN Architectures and NN Learning
2. Programming for Deep Learning.

SMAI-M20-L30:Neural Network Learning

C. V. Jawahar

IIIT Hyderabad

October 26, 2020

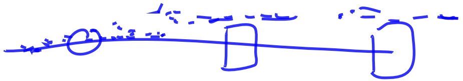
Announcements

- Quiz 2: Tentatively for next week. (say wed.)
- Refresh your programming and Jupyter notebooks.

Class Review

SVM and Kernels: Revisiting the illustrative problems we already solved.

- if samples scale in magntitude what happens to the solutions?
- if more samples come around a sample, what happens to the number of support vectors and α?
- ExOR and Paritty problems
- Does addition of new samples lead to new SVs?



Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs (hard margin, soft margin, kernel)
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Neural Network Architectures and Learning** (i) Neuron model, Single Layer Perceptrons (ii) SLP(iii) MLP. (iv) Backpropagation
- **Other Topics:**
  - Linear Regression
  - Probabilistic View, Bayesian View, MLE
  - Gradient Descent: Stochastic and Batch GD
  - Loss Functions and Optimization
  - Eigen Vector based optimization
  - Kernel Functions and Kernel Matrix

This Lecture:

1. Acctivations in Neural Networks
    1. Sigmoid and Tanh
    2. ReLU and L-ReLu
2. Learning in NNs
    1. Local minima and non-convexity
    2. Plateaus and saddle points
    3. Vanishing and Exploding Gradients

Questions? Comments?

Blank

Discussions Point - I

What happens to the NN Loss graphs on “Train” and “Val/Test” during the iterations?

1. How smooth is the graph?
2. When do we stop?
3. When is it overfitting?

Blank

Blank

Blank

Discussions Point -II

We know the functions like ReLu and Hinge Loss used in NN/ML.

- What makes these functions special for Gradient Descent (compared to functions like sigmoid or tanh)?

Blank

Blank

Discussion Point - III

(Advanced; Out of Syllabus!!) We know that a new kernel can be defined in terms of existing kernels:

$$\sum_{i=1}^{K} \alpha_i \kappa(\cdot,\cdot)$$

Then why don't we formulate the overall learning problem in SVM, including that of learning these α_i

1. Discuss why it is a good idea?
2. How do we use it for "fusing" different features?
3. Why do we limit to $\sum$?

See some of the works relevant[1] and [2]. Read later.

[1] http://manikvarma.org/pubs/varma07c.pdf

[2] https://cvit.iiit.ac.in/images/ConferencePapers/2009/Rakesh09More.pdf

Blank

Blank

Discussion Point - IV

(Advanced; Out of Syllabus!!)
We know that linear SVMs are superefficient (compared to K-SVMs).
Can we find a $\phi()$ corresponding to a Kernel and solve the problem as

$$\mathbf{w}^T \phi(x)$$

Indeed, this may become difficult for many kernels (eg. RBFs). **why?**
Can we find a finite dimensional approximation of $\phi()$? How does it help in speeding up SVM with no major reduction in accuracy?
read [3] and [4] later.

1. Discuss why it is a good idea?
2. Suggest an application where speed matters (eg. in the reference is that ofobject detection).

[3] https://cvit.iiit.ac.in/images/ConferencePapers/2010/Sreekanth10Generalized.pdf
[4] https://www.robots.ox.ac.uk/ vgg/publications/2011/Vedaldi11/vedaldi11.pdf

Blank

Blank

What Next:?

1. NN Architectures and NN Learning
2. Programming for Deep Learning.

1 Problem Set for In-class Problem Solving

Answer with sufficient details with insightful discussions (not just factual answers) and submit by 11am. (details discussed in the last two lecture sessions)

1. Consider a linearly separable data

$$(-1, +1), (0, -1), (+1, -1)$$

   (a) Write the primal problem. Geometrically show the feasible region in a w, b 2D plane. Show the optimal w as -2 and b as -1. Geometrically validate the solution.

   (b) Write the dual objective $J(\alpha_1, \alpha_2, \alpha_3)$. And the constraint $\sum_i \alpha_i y_i = 0$. Show that the solution is $\alpha_1 = \alpha_2 = 2$ and $\alpha_3 = 0$. Which are the support vectors then?

   (c) Show that the dual solution also leads to the same w and b.

2. Consider a linearly in-separable 1D data

$$(-1, +1), (0, -1), (+1, +1)$$

   Demonstrate how the problem can be solved with a quadratic kernel $(p^T q + 1)^2$

   (a) Write the dual objective with the constraint $\sum_i \alpha_i y_i = 0$

   (b) Substitute for α_2 from the constraint in the objective. (i.e., eliminate α_2)

   (c) Differentiate wrt α_1 and α_3 and equate to zero. Solve to obtain $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = 2$

   (d) For a new sample the decision is

$$sign(\sum_i \alpha_i y_i \kappa(x_i, x) + b)$$

   Simplify and get this expression as $2x^2 - 1$. (Assume $b = -1$. Indeed, you can find b yourself.)

3. Consider the ExOR Problem:

x_1	x_2	y
-1	-1	-1
-1	+1	+1
+1	-1	+1
+1	+1	-1

   (a) Consider a kernel $(p^T q + 1)^2$ and construct the Kernel matrix

   (b) Write the dual objective in terms of α_1, α_2, α_3 and α_4.

   (c) Differentiate wrt α_1, α_2, α_3 and α_4 and equate to zero. You get four equations.

   (d) Solve (guess) the values for α_1, α_2, α_3 and α_4. Same?

   (e) We know $\phi() = [1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2]^T$. Find w as $\sum_i \alpha_i y_i \phi(x_i)]$. Does it solve the exor? why?

4. VC dimension of a function class is the largest set that can be shattered by a member of the function class.

   (a) Interval $[\alpha_1, \alpha_2]$ on real line. (i.e., +ve is inside $[\alpha_1, \alpha_2]$ else negative)

   (b) Axis parallel rectangles in 2D. (i.e., positive inside and negative outside)

   (c) Convex Polygons in 2D (inside positive and outside negative)

   (d) Union of intervals in 1D.

5. Starting from the Lagrangian

$$L(\mathbf{w}, b, \alpha, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{N} \xi_i^2 - \sum_{i=1}^{N} \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i]$$

   Derive the dual function for L2 SVM

SMAI-M20-L28: Illustrative Numerical Problems (Cont.)

C. V. Jawahar

IIIT Hyderabad

October 21, 2020

Class Review

Backpropagation

1. Convergence
2. Relationship between Loss and Acccuracy
3. Accuracy on Train Set and Test Set
4. Characterizing Optimization Problem in MLP

Blank

Recap:

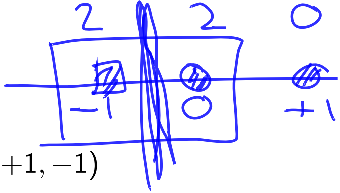
- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs (hard margin, soft margin, kernel)(vi) MLP (vii) BP
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - Probabilistic View, Bayesian View, MLE
    - Gradient Descent: Stochastic and Batch GD
    - Loss Functions and Optimization
    - Eigen Vector based optimization
    - Neuron model, Single Layer Perceptrons
    - Kernel Functions and Kernel Matrix

Blank

Illustrative Problem - I (Done: Q?)

Consider a linearly separable data

$$(-1, +1), (0, -1), (+1, -1)$$



$x = -1/2$

1. Write the primal problem. Geometrically show the feasible region in a w, b 2D plane. Show the optimal w as -2 and b as -1. Geometrically validate the solution. $wx + b = 1$
2. Write the dual objective $J(\alpha_1, \alpha_2, \alpha_3)$. And the constraint $\sum_i \alpha_i y_i = 0$. Show that the solution is $\alpha_1 = \alpha_2 = 2$ and $\alpha_3 = 0$. Which are the support vectors then?
3. Show that the dual solutiion also leads to the same w and b.

$w = \sum_{i} \alpha_i y_i x_i$ $\quad$ $\boxed{w = -2 \quad b = -1}$

Blank

Blank

Blank

Blank

Illustrative Problem -II (Done; Q?)

Consider a linearly in-separable 1D data

$$(-1,+1),(0,-1),(+1,+1)$$

Demonstrate how the problem can be solved with a quadratic kernel $(\mathrm{p}^T q + 1)^2$

1. Write the dual objective with the constraint $\sum_i \alpha_i y_i = 0$
2. Substitute for α_2 from the constraint in the objective. (i.e., eliminate α_2)
3. Differentiate wrt α_1 and α_3 and equate to zero. Solve to obtain $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = 2$
4. For a new sample the decision is

$$sign(\sum_i \alpha_i y_i \kappa(x_i, x) + b)$$

Simplify and get this expression as $2x^2 - 1$. (Assume $b = -1$. Indeed, you can find b yourself.)

Blank

Blank

Blank

Blank

Illustrative Problem - III

Consider the ExOR Problem:

x_1	x_2	y
-1	-1	-1
-1	+1	+1
+1	-1	+1
+1	+1	-1

1. Consider a kernel $(p^T q + 1)^2$ and construct the Kernel matrix
2. Write the dual objective in terms of α_1, α_2, α_3 and α_4.
3. Differentiate wrt α_1, α_2, α_3 and α_4 and equate to zero. You get four equations.
4. Solve (guess) the values for α_1, α_2, α_3 and α_4. Same?
5. We know $\phi() = [1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2]^T$. Find w as $\sum_i \alpha_i y_i \phi(x_i)]$. Does it solve the exor? why?

Blank

Blank

Blank

Blank

Blank

Blank

Illustrative Problem - IV

VC dimension of a function class is the largest set that can be shattered by a member of the function class.

1. Interval $[\alpha_1, \alpha_2]$ on real line. (i.e., $+$ve is inside $[\alpha_1, \alpha_2]$ else negative)
2. Axis parallel rectangles in 2D. (i.e., positive inside and negative outside)
3. Convex Polygons in 2D (inside positive and outside negative)
4. Union of intervals in 1D.

[1]

[1] VC Dimension is d, if a set of size d can be shattered completely and no set of size $d + 1$ can be shattered.

Blank

Blank

Blank

Blank

Blank

Blank

Discussion Point (Repeat, if time permits)

(Advanced; Out of Syllabus!!) We know that a new kernel can be defined in terms of existing kernels:

$$\sum_{i=1}^{K} \alpha_i \kappa(\cdot, \cdot)$$

Then why don't we formulate the overall learning problem in SVM, including that of learning these α_i

1. Discuss why it is a good idea?
2. How do we use it for "fusing" different features?
3. Why do we limit to $\sum$?

See some of the works relevant[2] and [3]. Read later.

[2] http://manikvarma.org/pubs/varma07c.pdf

[3] https://cvit.iiit.ac.in/images/ConferencePapers/2009/Rakesh09More.pdf

Blank

Blank

Discussion Point (Repeat, if time permits)

(Advanced; Out of Syllabus!!)
We know that linear SVMs are superefficient (compared to K-SVMs).
Can we find a $\phi()$ corresponding to a Kernel and solve the problem as

$$\mathbf{w}^T\phi(x)$$

Indeed, this may become difficult for many kernels (eg. RBFs). **why?**
Can we find a finite dimensional approximation of $\phi()$? How does it help in speeding up SVM with no major reduction in accuracy?
read [4] and [5] later.

1. Discuss why it is a good idea?
2. Suggest an application where speed matters (eg. in the reference is that ofobject detection).

[4] https://cvit.iiit.ac.in/images/ConferencePapers/2010/Sreekanth10Generalized.pdf

[5] https://www.robots.ox.ac.uk/ vgg/publications/2011/Vedaldi11/vedaldi11.pdf

Blank

Blank

What Next:?

Friday Lecture Session:

1. No formal online session
2. Submit the solution to the five problems we solved in class (11am on Friday). Write neatly with sufficient explanations and equations/pictures/sketches/details.

Next?

1. NN Architectures and NN Learning
2. Programming for Deep Learning.
3. Beyond Supervised Learning

SMAI-M20-L27: Illustrative Numerical Problems

C. V. Jawahar

IIIT Hyderabad

October 19, 2020

Class Review

MLP

1. Number of parameters
2. What can it solve?
3. What can it guarantee?

VC Dimension

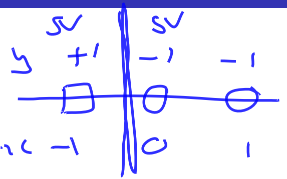
Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs (hard margin, soft margin, kernel)(vi) MLP
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - Probabilistic View, Bayesian View, MLE
    - Gradient Descent: Stochastic and Batch GD
    - Loss Functions and Optimization
    - Eigen Vector based optimization
    - Neuron model, Single Layer Perceptrons
    - Kernel Functions and Kernel Matrix

Blank

Illustrative Problem - I



Consider a linearly separable data

$x\ y\quad x_2\ y_2\quad x_3\ y_3$

$$(-1,+1),(0,-1),(+1,-1)$$

$wx+b=0$

1. Write the primal problem. Geometrically show the feasible region in a w, b 2D plane. Show the optimal w as -2 and b as -1. Geometrically validate the solution.
2. Write the dual objective $J(\alpha_1, \alpha_2, \alpha_3)$. And the constraint $\sum_i \alpha_i y_i = 0$. Show that the solution is $\alpha_1 = \alpha_2 = 2$ and $\alpha_3 = 0$. Which are the support vectors then?
3. Show that the dual solutiion also leads to the same w and b.

Blank

Blank

Blank

Blank

Blank

Illustrative Problem -II

Consider a linearly in-separable 1D data

$$(-1,+1),(0,-1),(+1,+1)$$

Demonstrate how the problem can be solved with a quadratic kernel $(\mathrm{p}^T q + 1)^2$

1. Write the dual objective with the constraint $\sum_i \alpha_i y_i = 0$
2. Substitute for α_2 from the constraint in the objective. (i.e., eliminate α_2)
3. Differentiate wrt α_1 and α_3 and equate to zero. Solve to obtain $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = 2$
4. For a new sample the decision is

$$sign(\sum_i \alpha_i y_i \kappa(x_i, x) + b)$$

   Simplify and get this expression as $2x^2 - 1$. (Assume $b = -1$. Indeed, you can find b yourself.)

Blank

Blank

Blank

Blank

Illustrative Problem - III

VC dimension of a function class is the largest set that can be shattered by a member of the function class.

1. Interval $[\alpha_1, \alpha_2]$ on real line. (i.e., $+$ve is inside $[\alpha_1, \alpha_2]$ else negative)
2. Axis parallel rectangles in 2D. (i.e., positive inside and negative outside)

1

[1] VC Dimension is d, if a set of size d can be shattered completely and no set of size $d + 1$ can be shattered.

Blank

Blank

Illustrative Problem - IV

Consider the ExOR Problem:

x_1	x_2	y
-1	-1	-1
-1	$+1$	$+1$
$+1$	-1	$+1$
$+1$	$+1$	-1

1. Consider a kernel $(p^T q + 1)^2$ and construct the Kernel matrix
2. Write the dual objective in terms of α_1, α_2, α_3 and α_4.
3. Differentiate wrt α_1, α_2, α_3 and α_4 and equate to zero. You get four equations.
4. Solve (guess) the values for α_1, α_2, α_3 and α_4. Same?
5. We know $\phi() = [1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2]^T$. Find w as $\sum_i \alpha_i y_i \phi(x_i)]$. Does it solve the exor? why?

Blank

Blank

Blank

Blank

Blank

Blank

What Next:? (next three)

1. NN Architectures and NN Learning
2. Programming for Deep Learning.
3. Beyond Supervised Learning

SMAI-M20-L26:Nonlinear methods: SVM, Kernels and MLP

C. V. Jawahar

IIIT Hyderabad

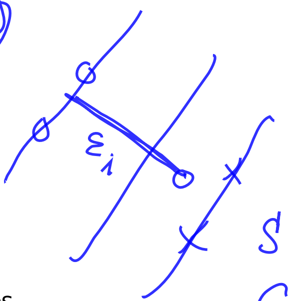
October 16, 2020

Class Review

"tight"

$\boxed{\geq \alpha}$

KKT

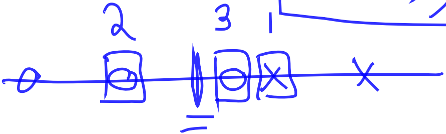


S mall

- Hard Margin SVM
- Soft Margin SVM
- Kernel SVM
- Specific Questions based on lecture notes.

Properties and limiting cases.

$\boxed{\geq 1 - \varepsilon_i}$



Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs (hard margin, soft margin, kernel)(vi) MLP
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - Probabilistic View, Bayesian View, MLE
    - Gradient Descent: Stochastic and Batch GD
    - Loss Functions and Optimization
    - Eigen Vector based optimization
    - Neuron model, Single Layer Perceptrons
    - Kernel Functions and Kernel Matrix

This Lecture:

1. MLP Architecture
   1. Role of Activations
   2. Regression, Classification and choice of output neurons.
   3. Expressive power of neural networks.
2. Chain rule for computing gradients
   1. How gradients can be computed
   2. What should we keep in mind while defining the layers.
3. Backpropagation through chain rule.
   1. Appreciate how BP works
   2. Why "back" in the BP ?
4. Kernel Ridge Regression
   1. Another example of Kernelization
   2. Familiarity of K and Φ

Questions? Comments?

Blank

Discussions Point - I

We have a three-class classification problem. We want to use an MLP

A we can use a single output neuron and force it to output 0, 1 and 2 corresponding to the three classes.

B we can have three output neurons with classes coded as [1,0,0], [0,1,0] and [0,0,1].

1. Which one will you prefer? Why?

Blank

Blank

Blank

Discussions Point -II

Kernel Ridge Regression:

1. We used the result

$$(BA + \lambda I)^{-1} B \ = B(AB + \lambda I)^{-1}$$

   verify this.

2. What are the steps in the training time for K-Ridge regression? What. are the steps during testing?

Blank

Blank

Discussion Point - III

(Advanced; Out of Syllabus!!) We know that a new kernel can be defined in terms of existing kernels:

$$\sum_{i=1}^{K} \alpha_i \kappa(\cdot,\cdot)$$

Then why don't we formulate the overall learning problem in SVM, including that of learning these α_i

1. Discuss why it is a good idea?
2. How do we use it for "fusing" different features?
3. Why do we limit to $\sum$?

See some of the works relevant[1] and [2]. Read later.

[1] http://manikvarma.org/pubs/varma07c.pdf

[2] https://cvit.iiit.ac.in/images/ConferencePapers/2009/Rakesh09More.pdf

Blank

Blank

Discussion Point - IV

(Advanced; Out of Syllabus!!)
We know that linear SVMs are superefficient (compared to K-SVMs).
Can we find a $\phi()$ corresponding to a Kernel and solve the problem as

$$\mathbf{w}^T\phi(x)$$

Indeed, this may become difficult for many kernels (eg. RBFs). **why?**
Can we find a finite dimensional approximation of $\phi()$? How does it help in speeding up SVM with no major reduction in accuracy?
read [3] and [4] later.

1. Discuss why it is a good idea?
2. Suggest an application where speed matters (eg. in the reference is that ofobject detection).

[3] https://cvit.iiit.ac.in/images/ConferencePapers/2010/Sreekanth10Generalized.pdf
[4] https://www.robots.ox.ac.uk/ vgg/publications/2011/Vedaldi11/vedaldi11.pdf

Blank

Blank

What Next:? (next three)

1. SVMs and Kernels (winding up?)
2. NN Architectures and NN Learning
3. Programming for Deep Learning.
4. Next Lecture (Mon or Wed): Problem Solving related to SVM and Kernels. (no new video content!).

SMAI-M20-L25: Nonlinear methods: SVM, Kernels and MLP

C. V. Jawahar

IIIT Hyderabad

October 12, 2020

Class Review

$$[\phi(x_1) \cdots \phi(x_d)] \qquad \boxed{x} \longrightarrow \boxed{\phi(x)}$$

Questions is based on the brief review of Kernels you had seen at: `https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0`(shared in the class last week).

(see specially Section 4 for today)

$$[x_1 \cdots x_N]$$

1. Relationship between Data Matrix and Kernel Matrix.
2. How computations such as (i) centering (ii) distance (iii) normalization can be done in the feature space ($\phi()$).

$$K(\quad) = [\phi(x) - \phi(y)]^T [\phi(x) - \phi(y)]$$

Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs (hard margin, soft margin, kernel)
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - Probabilistic View, Bayesian View, MLE
    - Gradient Descent: Stochastic and Batch GD
    - Loss Functions and Optimization
    - Eigen Vector based optimization
    - Neuron model, Single Layer Perceptrons
    - Kernel Functions and Kernel Matrix

This Lecture:

1. Kernel SVMs:
   - What happens during training?
   - What happens during testing?
2. MLP:
   - Introduction to a typical MLP architectures
   - Appreciating “Deep MLP” (aka Deep Neural Network) as feature transformation followed by a classification.

Questions? Comments?

Blank

Blank

Blank

Discussions Point -I

A typical SVM training leaves you with the set of Support Vectors (SV) and the corresponding αs (for all others αs are zero).

1. Comment: "The ratio $\frac{|SV|}{N}$ gives us an idea of the error rate." Why? Argue?
2. Is this an upper bound or lower bound of LOO?
3. If we change the kernel, does the SVs change?

Blank

Blank

Blank

Discussions Point -II

Starting from the Lagrangian

$$L(\mathbf{w}, b, \alpha, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i]$$

i.e., Derive the dual function for L2 SVM
(Write on a paper, complete and submit later.)

Blank

Blank

Blank

Blank

Discussion Point - III

An MLP has 3 inputs, two hidden layers each of 5 neurons, and three output neurons. No bias anywhere. Hidden neurons have sigmoid activations. Output neurons have linear activations.
How many parameters are there to learn?

What Next:? (next few)

1. SVMs and Kernels (winding up)
2. MLP and Backpropagation

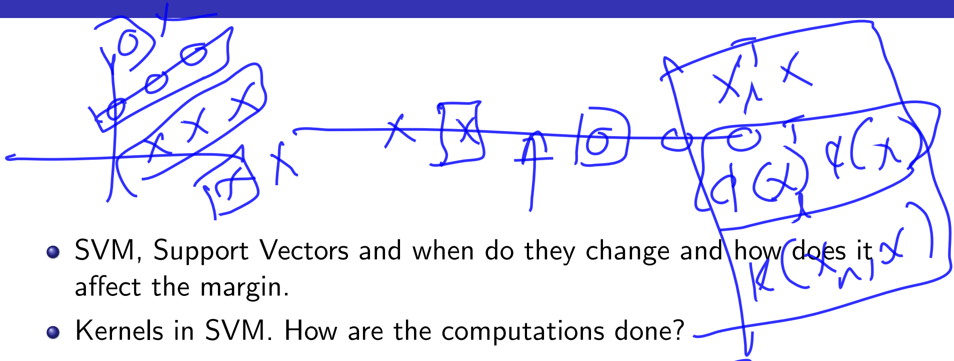
SMAI-M20-L24: SVM and Kernels

C. V. Jawahar

IIIT Hyderabad

October 9, 2020

Class Review



- SVM, Support Vectors and when do they change and how does it affect the margin.
- Kernels in SVM. How are the computations done?

all samples

$$\boxed{\sum_{i=1}^{N} \alpha_i}$$

$$\underline{w^T x}$$

$$\sum_{i=1}^{N} \alpha_i y_i x_i$$

Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - Probabilistic View, Bayesian View, MLE
    - Gradient Descent: Stochastic and Batch GD
    - Loss Functions and Optimization
    - Eigen Vector based optimization
    - Neuron model, Single Layer Perceptrons
    - Kernel Functions and Kernel Matrix

This Lecture:

1. Kernels:
   - Examples of PD Kernels (or valid kernels)
   - Kernels on structures (or non-vector) data
2. SVM:
   - Derivation of Dual from the Primal

Questions? Comments?

Blank

Discussions Point -I

We know the perceptron classification as

$$sign(\sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}))$$

and the kernel perceptron as

$$sign(\sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}))$$

1. Does the kernel perceptron yield a nonlinear boundary?
2. Assume the samples were in 2D, how do we plot (or visualize the decision boundary)?

Blank

Blank

Blank

Discussions Point -II

We had seen how that

$$\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

1. How do we find b^* for hard margin SVM?
2. How do we find b^* for soft margin SVM? [1]
3. Support Vectors (SV) are the vectors where α_is non-zero. Why α zero for non-support vector? [2]

[1] https://stats.stackexchange.com/questions/451868/calculating-the-value-of-b-in-an-svm

[2] https://stats.stackexchange.com/questions/355661/svm-why-alpha-for-non-support-vector-is-zero

Blank

Blank

Discussion Point - III

Starting from the Lagrangian

$$L(\mathbf{w}, b, \alpha, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{N}\xi^2 - \sum_{i=1}^{N}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 - \xi_i]$$

i.e., Derive the dual function for L2 SVM
(Write on a paper, complete and submit later.)

What Next:? (next few)

1. SVMs and Kernerls
2. MLP and Backpropagation

SMAI-M20-L23: SVMs and Kernels

C. V. Jawahar

IIIT Hyderabad

October 5, 2020

Announcements

- The number of students who watch and prepare for the sessions is low.
- A number of students are finding it hard to find time. Partly understandable in these times. But not fully acceptable.
- As we move forward, you could expect questions beyond MCQ. (in CR, Quiz, In class). Need to nurture the skill of solving problems on paper.
- Fine tuning the evaluation scheme.

Item	Orig.	Updated
Class Review (⇑)	10	25
Home Works (⇓)	40	30
Quiz	25	25
Assignments (⇓) Problem Solving (+)	25	20

Class Review

$$\underline{\underline{z}} = \kappa(p, q)$$

We know the kernel $\kappa()$ and the feature map $\phi()$.
Let us start with samples in 2D and

- Understand how $\phi()$ and $\kappa()$ are related in many specific cases.
- Is it unique?

$$\kappa(p, q) = \phi(p)^T \phi(q)$$

Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures (v) SVMs
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - Probabilistic View, Bayesian View, MLE
    - Gradient Descent: Stochastic and Batch GD
    - Loss Functions and Optimization
    - Eigen Vector based optimization
    - Neuron model, Single Layer Perceptrons
    - Kernel Functions and Kernel Matrix

This Lecture:

1. Soft Margin SVMs
    - SVM as a classifier that maximizes the margin.
    - How do we make the constraints “soft”.
2. Decision Tree Classifier
    - Popular. Simple. Interpretable.
    - Recursive Design. Node test.
3. Kernel Perceptron
    - Illustrative: Kernalizing a linear algorithm.

Questions? Comments?

Blank

Blank

Discussions Point - I

Consider there are 100 samples in 6 dimensions (i.e., $N = 100$ and $d = 6$) and a binary classification (50 each in class + and class -) (i.e., (50+,50-) If we use ith (1 to 6) feature based the node-test at the root, the two subsets formed are as

A (25+,25-) and (25+,25-)

B (45+,35-) and (5+,15-)

C (25+,2-) and (25+,48-)

D (40+,10-) and (10+,40-)

E (0+,50-) and (50+,0-)

F (0+,0-) and (50+,50-)

1. What do you prefer as the node-test? (list the options in the decreasing order of your preference)
2. "When $d = 6$, there can be only six possible splits" True or False?
3. Does your decision of node test "guarantee" that this is the best choice?

Information Gain

In decision tree [1] design, a popular way to do this is byestimating the information gain[2]
Basic Idea: “Estimate the entropy of the parent set. Estimate the entropy of the children sets (with different attributes) and select the best that removes most uncertainty.”

[1] https://en.wikipedia.org/wiki/Decision$_{t}ree_{l}earning$

[2] https://en.wikipedia.org/wiki/Information$_{g}ain_{i}n_{d}ecision_{t}rees$

Blank

Blank

Blank

Discussion Point - II

We know the softmargin SVM problem as

$$Min\ \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i^2$$

subject to:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i\ \ \forall i$$

$$\xi_i \geq 0$$

1. Why $\xi_i \geq 0$? required in the constraint? Why ξ_i in the objective?
2. "If a specific problem has a hard margin possible, ξ_i will all be zero"? True or False?
3. If C is very small (say $C = 0$), what does it mean? what do you expect to see in the final solution?
4. If C is very large (say $+\infty$), what does it mean? what do you expect to see in the final solution?
5. How do we choose C?

Blank

Blank

Blank

Blank

Blank

Discussions Point -III

We know the perceptron classification as

$$sign(\sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}))$$

and the kernel perceptron as

$$sign(\sum_{i=1} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}))$$

1. Does the kernel perceptron yield a nonlinear boundary?
2. Assume the samples were in 2D, how do we plot (or visualize the decision boundary)?

Blank

Blank

Blank

What Next:?

1. **More on SVMs and Kernels**

SMAI-M20-L22: Introduction to SVM

C. V. Jawahar

IIIT Hyderabad

September 30, 2020

Announcements

- **Quiz 1**
  - The same time as last week.
  - The same set of topics we planned.
  - All objective. (similar to Class Review)

Class Review

In the context of binary classification and LDA:

- What do we know about the direction of the discriminant vector?
- What do we know about the objective?
- What do we know bout the uniqueness of the solution?
- What do we know about the ranks of S_w and S_B?
- How does singularity of S_B and S_w affect us?

Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - MLE
    - Gradient Descent
    - Stochastic and Batch GD
    - Eigen Vector based optimization
    - Neuron model
    - Loss Functions and Optimization
    - Kernel Functions and Kernel Matrix

This Lecture:

1. Introduction to SVMs
   - SVM as a classifier that maximizes the margin.
2. Solving Logistic Regression as GD
   - From objective to GD and Regularization
3. LDA: Extending to Multi-Class

Questions? Comments?

Blank

Blank

Discussions Point - I

Consider Five 1D samples:

$$(1,+),\ (2,+),\ (7,-),\ (6,-),\ (11,-)$$

1. Which will be a valid decision boundary with perceptron criteria? (i) 2.5 (ii) 6.5 (ii) 5.00 (iv) 4.00
2. What will be the optimal decision boundary with SVM criteria? (i) 2.5 (ii) 6.5 (ii) 5.00 (iv) 4.00
3. Assume we add a sample $(8,-)$ to the training set, will SVM decision boundary change? why? what is the new one?
4. Assume we add a sample $(4,-)$ to the training set, will SVM decision boundary change? why? what is the new one?

Blank

Blank

Blank

Discussions Point -II

We know the objective of logistic regression as:

$$\sum_{i=1}^{N} y_i \log(g(\mathbf{w}^T \mathbf{x}) + (1 - y_i) \log(1 - g(\mathbf{w}^T \mathbf{x})$$

Derive the gradient ascent update equation

Hint:

$$w^{k+1} = w^k + \eta \sum_{i=1}^{N} (y_i - g(\mathbf{w}^T \mathbf{x}_i))\mathbf{x}_i$$

Blank

Blank

Blank

Blank

Blank

Discussion Point - III

How do you compare the GD rule with that of if we had used an MSE loss between predicted probabilities and our actual labels (probabilities)?

$$w^{k+1} = w^k + \eta \sum_{i=1}^{N} (y_i - g(\mathbf{w}^T \mathbf{x}_i))\mathbf{x}_i$$

Blank

Blank

Blank

What Next:? (next three)

1. Winding up (i) Logistic Regression (ii) Multi-Class Classification and (iii) LDA
2. **SVMs and Kernels**

SMAI-M20-L21: Introduction to Kernels

C. V. Jawahar

IIIT Hyderabad

September 28, 2020

Announcements

- **Quiz 1**
  - The same time as last week.
  - The same set of topics we planned.
- **Course Evaluation**
  - We may have to minorly tweak the course evaluation models.
  - Will be announced next week.

Class Review

We are given a problem of Multi-Class Classification with

- DDAG
- BHC
- One-vs-Rest.

Now we want to know:

1. What is the computational advantages of each architecture?
2. How many nodes (binary classifiers) are required?
3. When is the decision ambiguous?
4. How do we compute the accuracy of the system from accuracy of the individual classifiers?

Blank

Recap:

- **Supervised Learning:** Formulation, Conceptual Issues, Concerns etc.
- **Classifiers:** (i) Nearest Neighbour, (ii) Notion of a Linear Classifier (iii) Perceptrons (iv) Bayesian Optimal Classifier (v) Logistic Regression (vi) Multiclass classification architectures
- **Dimensionality Reduction and Applications:** (i) Feature Selection and Extraction (ii) PCA (iii) LDA (iv) Eigen face
- **Matrix Factorization and Applications:** (i) SVD, (ii) Eigen Decomposition (iii) Matrix Completion (iv) LSI (v) Recommendations
- **Other Topics:**
    - Linear Regression
    - MLE
    - Gradient Descent
    - Stochastic and Batch GD
    - Eigen Vector based optimization
    - Neuron model
    - Loss Functions and Optimization

This Lecture:

1. Introduction to Kernels
   - Kernel Trick: A method of solving nonlinear problems with linear algorithms.
   - Kernel Function: $\kappa(p, q) = \phi(p)^T \phi(q)$
2. Extending the idea of Logistic Regression to Multi-Class
   - Classifiers output a score that a decision. Making fusion simpler.
   - Softmax: Find the maximum, normalize and probabilistic interpretation.

Questions? Comments?

Blank

Blank

Discussions Point - I

We had seen the Kernel

$$\kappa(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p})^T \phi(\mathbf{q})$$

with

$$\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q})^2$$

This means:

$$\phi : \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \rightarrow \begin{bmatrix} p_1^2 \\ p_2^2 \\ \sqrt{2} p_1 p_2 \end{bmatrix}$$

Is the feature map unique given the kernel?

Blank

Blank

Blank

Discussions Point -II

We now the $\phi()$ correspond to:

$$\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q})^2$$

What is the $\phi()$ correspond to:

$$\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2$$

(assume $\mathbf{p} \in R^2$)

Blank

Blank

Blank

Discussion Point - III

Assume there are N samples.
A kernel matrix $\mathbf{K}$ is defined as a matrix with (i, j)th element as the kernel computed with the i th and j th sample. i.e.,

$$K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$$

1. What is the dimension of the Kernel Matrix?
2. Is Kernel matrix square?
3. Is Kernel matrix symmetric?
4. Is Kernel matrix PSD?

Blank

What Next:? (next three)

1. Winding up (i) Logistic Regression (ii) Multi-Class Classification and (iii) LDA
2. SVMs and Kernels

SMAI-M20-L20: LDA

C. V. Jawahar

IIIT Hyderabad

September 25, 2020

Announcements

1. **Quiz 1** on Next Wed.
2. **Topics:** Topics remain the same.
3. Most-Likely the same time.
4. Any other announcements: by Monday.

Class Review

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^{k} + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve.
Given $\mathbf{w}^0$. What do we know about $\mathbf{w}^1$ and $\mathbf{w}^2$?

Blank

Recap:

- Supervised Learning:
  - Notions of Training, Validation and Testing; Loss Function and Optimization, Generalization, Overfitting, Occam's razor, Model Complexity, Bias and Variance, Regularization.
  - Performance Metrics, Estimating error using validation set.
- Approaches:
  - Optimal Decision as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2, MLE
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Application of PCA: Eigen Face
  - Matrix Factorization for Data Matrices (SVD, Eigen Decomposition)
  - Application of Matrix Factorization: LSI, Matrix Completion, Recommendation Systems)
  - Nearest Neighbour, Linear Discriminants
  - Gradient Descent
  - Linear Regression: Closed form, GD, Regularization, Optimization
  - Perceptron Algorithm and Neuron Model
  - Logistic Regression
  - LDA
  - Multi-Class Classification Architectures

This Lecture:

1. **Logistic Regression - III**
   - Insight into LR objective
2. **LDA - II**
   - LDA solution
3. **Multi-Class Classification - II**
   - D-DAG

Questions? Comments?

Blank

Discussions Point - I

We know the solution to LDA as

$$\mathbf{w}^* = \alpha \mathbf{S}_W^{-1}[\mu_{\mathbf{A}} - \mu_{\mathbf{B}}]$$

A potential worry is "If $\mathbf{S_w}$ is singular? "

- Suggest a configuration of the data when $\mathbf{S}_w$ can be singular?
- Suggest solutions to handling this singularity problem while computing $\mathbf{w}^*$?

Blank

Blank

Blank

Discussions Point -II

Are there any design considerations in D-DAG?

We know the DAG way of arranging pair-wise classifiers.

(Assume we have 4 classes and 6 pairwise classifiers. We had seen in the micro-lecture.)

What will you prefer as the root classifier?

- the highest accuracy classifier
- the least accuracy classifier
- any classifier. This does not matter in the design.

Any other insight into how the classifiers/class to be arranged?[1]

[1] Read later: An old but relevant analysis: https://cvit.iiit.ac.in/images/ConferencePapers/2003/pavan03multiclass.pdf

Blank

Blank

Discussion Point - III

Comment on the following three different ways of formulating the loss for a binary classification[2]:

1. $\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$
2. $\sum_{i=1}^{N}(y_i - g(\mathbf{w}^T\mathbf{x}_i))^2$
3. LR objective

[2] Read later: somewhat relevant reference: http://books.jackon.me/Cross-Entropy-vs-Squared-Error-Training-a-Theoretical-and-Experimental-Comparison.pdf

Blank

What Next:? (next three or even more)

1. More on LR, Multi-Class Classification, Dimensionality Reduction
2. Intro to SVMs and Kernels.

SMAI-M20-L19: Logistic Regression

C. V. Jawahar

IIIT Hyderabad

September 23, 2020

Class Review

What weights can lead to various logic gates in a Single Layer Perceptron?

- with $\{-1, +1\}$logic or $\{0, 1\}$ logic
- AND, OR, NAND, NOR, NOT etc.
- with various definitions of "activations" $\phi()$.

Blank

Announcements

Quiz 1

- Mostly based on topics upto (including PCA). No focus on perceptron, GD, Logistic Regression.
- Assumes. you followed lecture videos, class reviews, discussion points, home works.
- Four parts (each like a class review). Each of 5 questions. Total: 20 questions. Modular.
  - No discussions with friends, experts, classmates during the quiz.
  - Not a quiz of 20 question. Think of it as four class reviews.
  - One part is numerical. Keep a calculator, if required.
  - One part may ask you to refer to the class material. Keep course material, your notes, a computer accessible. Keep a pen/paper handy.
  - Each part: 10 mins within a span of 15 mins (say Part 1 during 6.30-6.45pm). Take care of brief network/power outages. Brief brakes required by individuals.
  - Any issues, email course email address. (not to individuals). Any clarifications/announcements will be in the "Quiz Channel".
  - People who miss quiz due to any genuine issue, do not attempt. Write email. We will see how to take care, including a possible extra one.

Recap:

- Supervised Learning:
  - Notions of Training, Validation and Testing; Loss Function and Optimization, Generalization, Overfitting, Occam's razor, Model Complexity, Bias and Variance, Regularization.
  - Performance Metrics, Estimating error using validation set.
- Approaches:
  - Optimal Decision as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2, MLE
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Application of PCA: Eigen Face
  - Matrix Factorization for Data Matrices (SVD, Eigen Decomposition)
  - Application of Matrix Factorization: LSI, Matrix Completion, Recommendation Systems)
  - Nearest Neighbour, Linear Discriminants
  - Gradient Descent
  - Linear Regression: Closed form, GD, Regularization, Optimization
  - Perceptron Algorithm and Neuron Model
  - Logistic Regression

This Lecture:

1. **Logistic Regression - II**
    - How do we formulate the Logistic Regression Objective as MLE?
2. **LDA/ Fisher**
    - Linear Discriminant Analysis or Fisher Discriminant
    - Supervised Dimensionality Reduction
3. **Multi Class Classification**
    - Why it is non-trivial for multi-class?
    - Practical issues (i) computational complexity (ii) Fusion/Decision making scheme

Questions? Comments?

Blank

Discussions Point - I

We know that LDA aims to:

- maximize between class variance
- minimize within class variance

Is it true to say that direction of PCA is orthogonal to that of. PCA?

Blank

Blank

Blank

Discussions Point -II

Comment on the following three ways of designing multi-class classifiers (number of classes = K):

1. One vs Rest Classifiers and some fusion scheme.
2. Pairwise Classification with Majority Voting
3. Binary Hierarchical Classification

based on the following dimensions:

- A Number of classifiers to be trained
- B Number of classifiers to be evaluated for testing a single sample.
- C Difficulty of the classification problem that each classifier solves (if required, assume classes are Multivariate Gaussian)

Comments A1, A2, A3, B1, B2, B3, C1, C2, C3 and finally what do you prefer? why? when?

Blank

Blank

Blank

What Next:? (next three)

1. Logistic Regression
2. Multi Class Classification (beyond binary)
3. More Dimensionality Reduction Schemes (eg. LDA/Fisher)