10707
Deep Learning

Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu

Midterm review

Midterm Review

- Polynomial curve fitting – generalization, overfitting
- Loss functions for regression

$$\mathbb{E}[L] = \int\int \left(t - y(\mathbf{x})\right)^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

- Generalization / Overfitting
- Statistical Decision Theory

Midterm Review

- Bernoulli, Multinomial random variables (mean, variances)
- Multivariate Gaussian distribution (form, mean, covariance)
- Maximum likelihood estimation for these distributions.
- Linear basis function models / maximum likelihood and least squares:

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta)$$

$$= -\frac{\beta}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\right)^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

Midterm Review

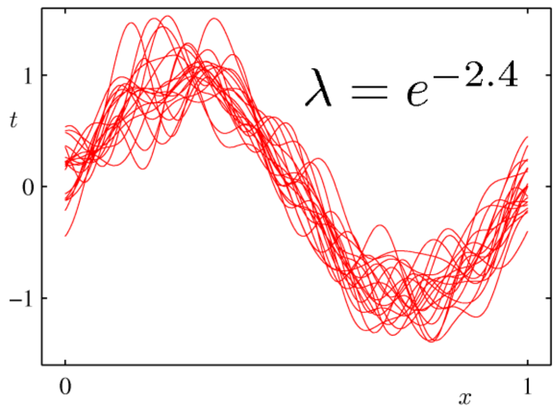
Ridge regression

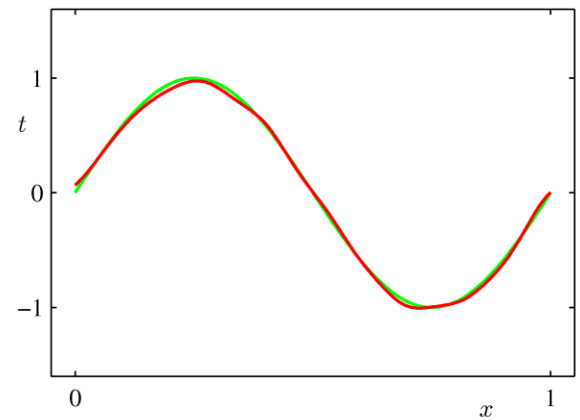
- Regularized least squares:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \qquad \mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

- Bias-variance decomposition.

High variance





Low bias

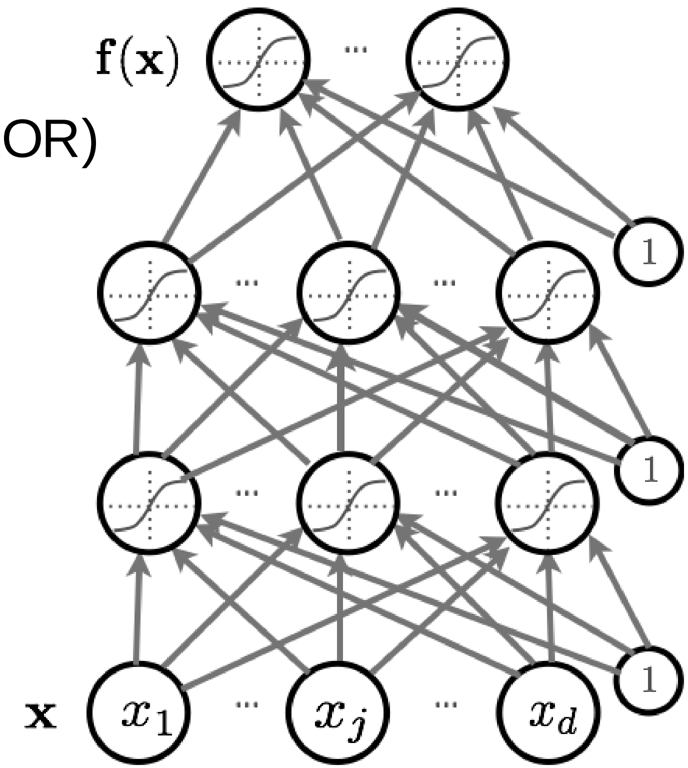
- Gradient Descend, SGD, Parameter Update Rules

Neural Networks

- How neural networks predict f(x) given an input x:
  - Forward propagation
  - Types of units
  - Capacity of neural networks (AND, OR, XOR)

- How to train neural nets:
  - Loss function
  - Backpropagation with gradient descent

- More recent techniques:
  - Dropout
  - Batch normalization
  - Unsupervised Pre-training



Neural Networks

- SGD Training, cross entropy loss, squared loss, ReLU activations
- Classification and regression with neural networks
- Regularization, Dropout, Batchnorm
- Forward Propagation and Backprop (computing derivatives)

- I may ask you to derive backprop for a regression / classification net with a single hidden layer, ask about what dropout and batchnorm are doing.

Conv Nets

- Convolutional networks leverage these ideas
  - Local connectivity
  - Parameter sharing
  - Convolution
  - Pooling / subsampling hidden units
  - Understanding Receptive Fields

  - I may give you a convnet and ask about its feedforward pass, computations

Graphical Models

- Directed and Undirected Graphs
  - ➢ Definition
  - ➢ Factorization Properties
  - ➢ Markov Blanket / Conditional Independence Properties
  - ➢ Gaussian Examples / Chain Graphs

  - ➢ I may give you graphical model and about conditional independence properties

RBMs

- Restricted Boltzmann Machines

  - ➢ Probably distribution, energy definition
  - ➢ Factorization Properties, Conditional probabilities
  - ➢ Maximum likelihood estimation (positive and negative phases)
  - ➢ Gradients estimation / derivation
  - ➢ Contrastive Divergence (CD) learning, Gibbs sampling

  - ➢ I may ask you to derive gradients of the loss function for learning

Deep Belief Networks

- DBNs, definition
  - Probably distribution, energy definition
  - Factorization Properties, Conditional probabilities
  - Greedy pretraining algorithm
  - Gradients estimation / derivation
  - Variational bound derivation

  - I may ask you for a definition of DBN, deriving variational bound for learning.