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

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

- 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:

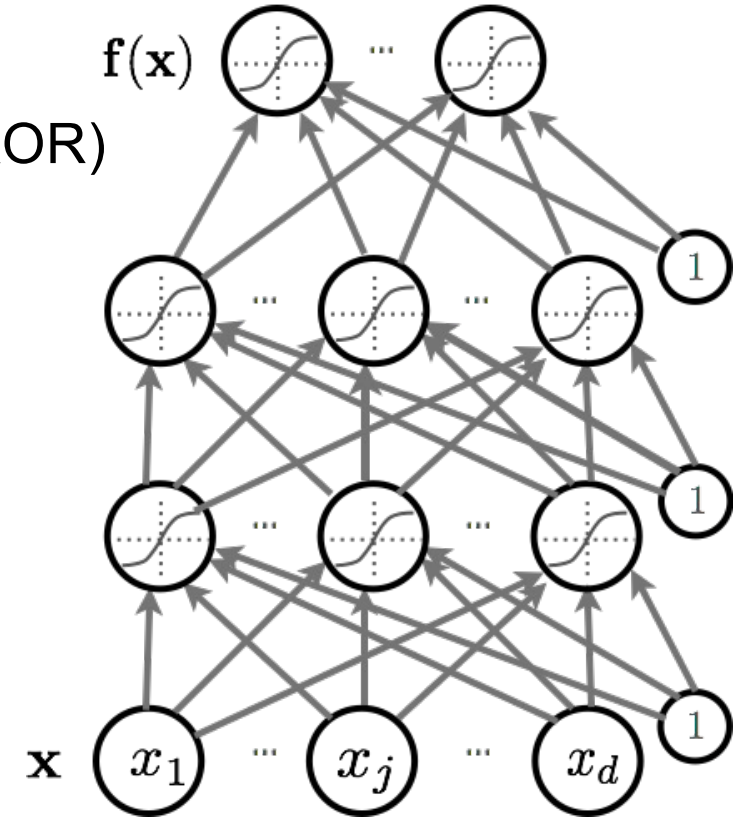
$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta)$$

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

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

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, Leaky ReLU activations
- Classification and regression with neural networks
- Regularization, Dropout, Batchnorm
- Forward Propagation and Backprop (computing derivatives)

- I may ask you about gradient updates of deep networks and some questions about differences between leaky ReLU and sigmoid activation functions.
- I may ask about what dropout and batchnorm are doing (eg how batchnorm works, and what differences are between training and test.

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, Deep Belief Networks, Autoencoders

- 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, and ask you to derive variational bound for learning, or show that the difference between log prob and its variation found is the KL divergence.