CSC411: Final Review

Shengyang Sun [2]

April 2, 2019

[2] Slides adapted from James Lucas & David Madras

Agenda

1. A brief overview
2. Some sample questions

Basic ML Terminology

The final exam will be on the entire course; however, it will be more heavily weighted towards post-midterm material. For pre-midterm material, refer to the midterm review slides on the course website.

- Feed-forward Neural Network (NN)
- Activation Function
- Backpropagation
- Fully-connected vs. convolutional NN
- Dimensionality Reduction
- Principal Component Analysis (PCA)
- Autoencoder
- Generative vs. Discriminative Classifiers
- Naive Bayes
- Bayesian parameter estimation
- Prior/posterior distributions
- Gaussian Discriminant Analysis (GDA)

Basic ML Terminology

The final exam will be on the entire course; however, it will be more heavily weighted towards post-midterm material. For pre-midterm material, refer to the midterm review slides on the course website.

- K-Means (hard and soft)
- Latent variable/factor models
- Clustering
- Gaussian Mixture Model (GMM)
- Expectation-Maximization (EM) algorithm
- Jensen's Inequality
- Matrix factorization
- Matrix completion
- Gaussian Processes
- Kernel trick
- Reinforcement learning
- States/actions/rewards
- Exploration/exploitation

Some Questions

Question 1

True or False:

1. PCA always uses an invertible linear map
2. K-Means will always find the global minimum
3. Naive Bayes assumes that all features are independent

Some Questions

Question 1

True or False:

1. PCA always uses an invertible linear map *False*
2. K-Means will always find the global minimum *False*
3. Naive Bayes assumes that all features are independent *False*

Some Questions

Question 1

True or False:

1. PCA always uses an invertible linear map *False*
2. K-Means will always find the global minimum *False*
3. Naive Bayes assumes that all features are independent *False*

Question 2

1. How can a generative model $p(\mathbf{x}|y)$ be used as a classifier?

Some Questions

Question 1

True or False:

1. PCA always uses an invertible linear map *False*
2. K-Means will always find the global minimum *False*
3. Naive Bayes assumes that all features are independent *False*

Question 2

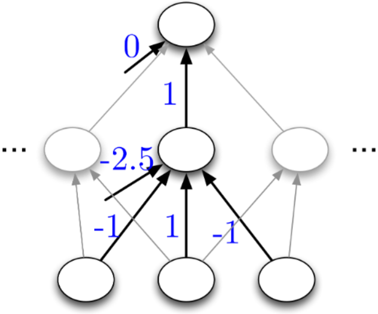
1. How can a generative model $p(\mathbf{x}|y)$ be used as a classifier?
2. Give one advantage of Bayesian linear regression over ML linear regression. Give a disadvantage.

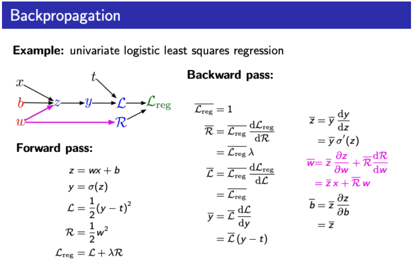
Subject Areas

1. Neural Networks
2. PCA
3. Probabalistic Models
4. Latent Variable Models
5. Bayesian Learning
6. Reinforcement Learning

Neural Networks

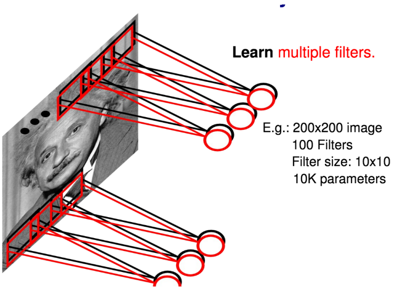
1. Forwarding given weights and biases
2. Why nonlinear activations are necessary?
3. Expressive power of neural networks
4. Backpropagation & gradient descent





Convolutional Neural Networks

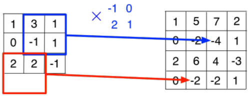
1. CNN architecture (kernels, channels, connections)
2. Local connection, weight sharing, pooling
3. How to perform convolutions ?
4. What specific functionality for some kernel ?



2-D convolution is analogous:

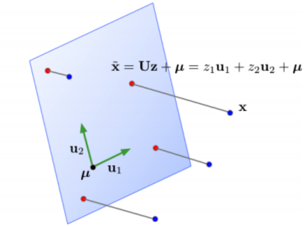
$$(A * B)_{ij} = \sum_s \sum_t A_{st} B_{i-s,j-t}.$$

$$\begin{bmatrix}1 & 3 & 1\\ 0 & -1 & 1\\ 2 & 2 & -1\end{bmatrix} * \begin{bmatrix}1 & 2\\ 0 & -1\end{bmatrix}$$



Principal Component Analysis (PCA)

1. Why dimensionality reduction ?
2. What does PCA reconstruction minimize?
3. How to perform PCA ?
4. What is the optimal PCA subspace given empirical Σ ?
5. Linear & non-Linear Autoencoders



$$\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \boldsymbol{\mu})$$

- In machine learning, $\tilde{\mathbf{x}}$ is also called the reconstruction of $\mathbf{x}$.
- $\mathbf{z}$ is its representation, or code.

Probabilistic Models

1. i.i.d.
2. Maximium Likelihood Estimation (MLE)
3. Generative $p(\mathbf{x}|y)$ vs. discriminative $p(y|\mathbf{x})$ classification

- Assume they're drawn from a Gaussian distribution with known standard deviation $\sigma = 5$, and we want to find the mean μ.
- Log-likelihood function:

$$\begin{aligned}\ell(\mu) &= \log \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2} \right) \right] \\ &= \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2} \right) \right] \\ &= \sum_{i=1}^{N} \underbrace{-\frac{1}{2}\log 2\pi - \log\sigma}_{\text{constant}} - \frac{(x^{(i)}-\mu)^2}{2\sigma^2}\end{aligned}$$

$$0 = \frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^{N} x^{(i)} - \mu$$

$$0 = \frac{\partial \ell}{\partial \sigma} = \frac{\partial}{\partial \sigma} \left[\sum_{i=1}^{N} -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2\sigma^2}(x^{(i)}-\mu)^2 \right]$$

Probabilistic Models (continued)

1. prior, likelihood, posterior. Bayes' rule

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta)$$

2. Bayesian parameter estimation
3. Maximum A Posteriori (MAP)

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \log p(\theta|\mathcal{D}) = \arg\max_{\theta} \log p(\mathcal{D}|\theta) + \log p(\theta)$$

Probabilistic Models (continued)

1. Gaussian Discriminant Analysis (mean, covariance)

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

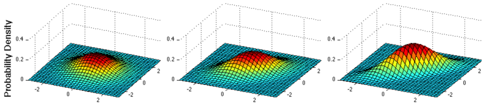


Figure: Probability density function

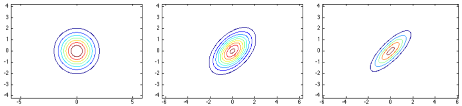


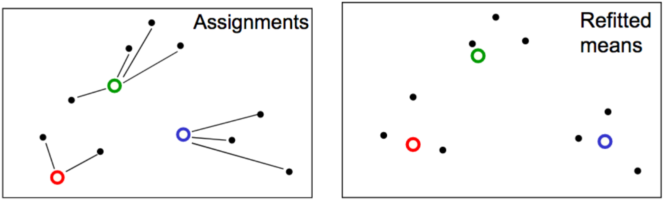
Figure: Contour plot of the pdf

2. Naive bayes: Assumes features independent **given the class.**

$$p(\mathbf{x}|t = k) = \prod_{i=1}^{d} p(x_i|t = k)$$

K-Means

1. Initialization, assigment, refitting
2. Convergence
3. Soft vs. hard K-means

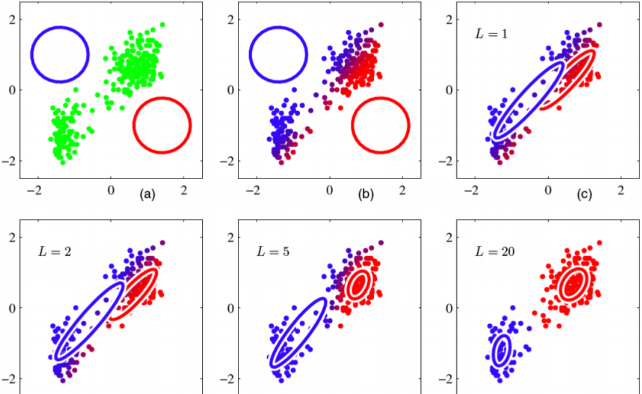


Gaussian Mixture Model (GMM)

1. latent (hidden) variables z,

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(z=k)p(\mathbf{x}|z=k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

2. E-Step: Compute the posterior over z given our current model
3. M-Step: Given z assignment, optimizes model parameters.



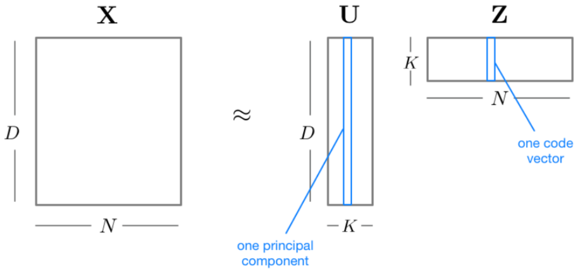
Expectation-Maximization (EM)

1. Why use latent variables ?
2. The EM lower bound

$$\sum_{n=1}^{N} \log p(\mathbf{x}^n; \theta) \geq \sum_{n=1}^{n} \mathbb{E}_{q_n(z^n)} \left[\log \frac{p(z^n, \mathbf{x}^n; \theta)}{q_n(z^n)} \right]$$

3. When is lower bound tight ?
4. E-Step and M-Step for optimizing the objective.

Matrix Factorization



1. Matrix Factorization: Rank-k approximation
2. Matrix completion . Alternating Least Squares (EM)
3. How K-Means can be seen as matrix factorization ?
4. Sparse Coding

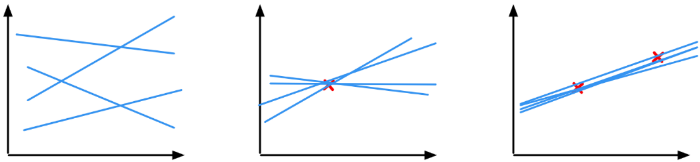
Bayesian Linear Regression

1. How can uncertainty in the predictions help us ?
2. Prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{S})$; Likelihood: $t|\mathbf{x}, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\top \psi(\mathbf{x}), \sigma^2)$
3. Posterior distribution $\mathbf{w}|\mathcal{D} \sim \mathcal{N}(\mu, \Sigma)$

$$\mu = \sigma^{-2} \Sigma \Psi^\top \mathbf{t}$$

$$\Sigma^{-1} = \sigma^{-2} \Psi^\top \Psi + \mathbf{S}^{-1}$$

4. Bayesian optimization, acquisition function

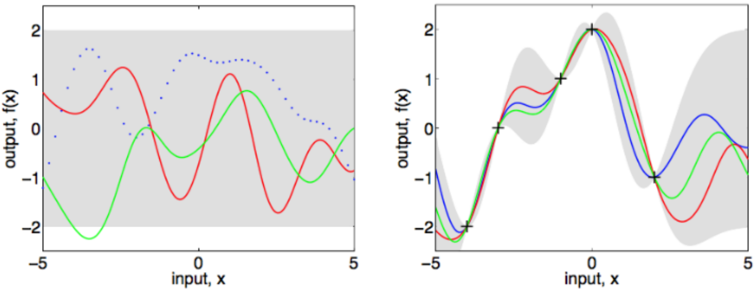


no observations one observation two observations

Gaussian Processes

1. Distribution over functions!
2. What requirement does the kernel $k(\cdot,\cdot)$ need to fullfill ?
3. How the kernel trick builds up the connection between kernel function and feature space ?

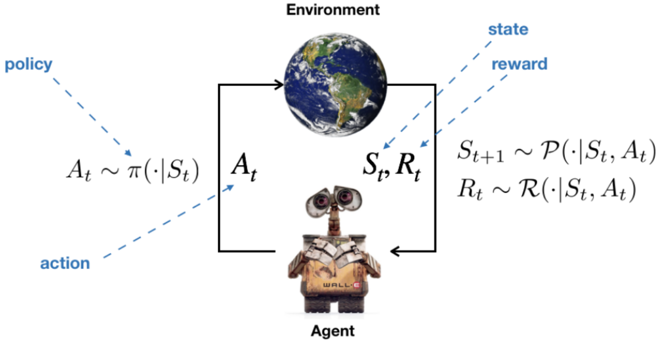
$$k(\mathbf{x},\mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$$



(a), prior (b), posterior

Reinforcement Learning

1. Choosing actions to maximize long-term reward
2. States, actions, rewards, policies, transition probability
3. Value function, Bellman Equation, value iteration
4. Exploration vs. Exploitation



Sample Question 1

Consider a 2-layer neural network, f, with 10-100-100 units in each layer respectively. We denote the weights of the network as $W^{(1)}$ and $W^{(2)}$.

a) What are the dimensions of $W^{(1)}$ and $W^{(2)}$? How many trainable parameters are in the neural network (ignoring biases)?

We will now replace the weights of f with a simple *Hypernetwork*. The Hypernetwork, h, will be a two layer network with 10 input units, 10 hidden units, and K output units where K is equal to the total number of trainable parameters in f. In each forward pass, the output of h will be reshaped and used as the weights of f.

b) How many parameters does h have (ignoring biases)?

c) How might we change the output layer to reduce the number of parameters? State how many trainable parameters h has with your suggested method. (HINT: use matrix factorization)

Q1 Solution

a) $W^{(1)} \in \mathbb{R}^{10\times 100}$, $W^{(1)} \in \mathbb{R}^{100\times 100}$. Total parameters: $10 \times 100 + 100 \times 100 = 11000$.

b) Total parameters: $10 \times 10 + 10 \times 11000 = 110100$.

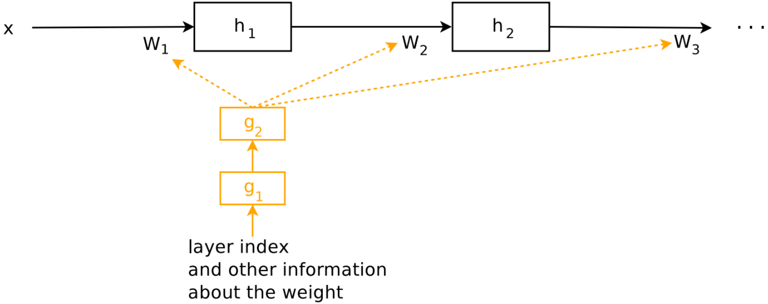
c) Output low rank approximations to each weight matrix. Instead of outputting $W^{(l)}$, output $U^{(l)}$ and $V^{(l)}$ such that $W^{(l)} \approx U^{(l)} V^{(l)}$. For example:

$$U^{(1)} \in \mathbb{R}^{10\times 2} \quad V^{(1)} \in \mathbb{R}^{2\times 100} \quad U^{(2)} \in \mathbb{R}^{100\times 2} \quad V^{(2)} \in \mathbb{R}^{2\times 100}$$

Now the total number of parameters is:
$10 \times 10 + 10 \times (2 \times 10 + 2 \times 100 + 100 \times 2 + 2 \times 100) = 6300$

Quick interlude: Hypernetworks

This isn't quite how Hypernetworks typically work...



See Ha et al. 2016 for details

Sample Question 2

a) State what conditions a function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ must satisfy to be a valid kernel function.

b) Prove that a symmetric matrix $K \in \mathbb{R}^{d \times d}$ is positive semidefinite if and only if for all vectors $\mathbf{c} \in \mathbb{R}^d$ we have $\mathbf{c}^T K \mathbf{c} \geq 0$.

Q2 Solution

a) Its Gram matrix, given by $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ must be positive semidefinite for any choices of $\mathbf{x}_1, \ldots, \mathbf{x}_d$.

b) First $\Rightarrow$: If K is PSD then there exists an orthonormal basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ with non-negative eigenvalues $\lambda_1, \ldots, \lambda_d$. We can write any vector $\mathbf{c}$ in this basis: $\mathbf{c} = \sum_{i=1}^{d} a_i \mathbf{v}_i$. Then,

$$\mathbf{c}^T K \mathbf{c} = (\sum_{i=1}^{d} a_i \mathbf{v}_i)^T K (\sum_{i=1}^{d} a_i \mathbf{v}_i) = \sum_{i=1}^{d} a_i a_j \mathbf{v}_i^T K \mathbf{v}_j = \sum_{i=1}^{d} a_i a_j \mathbf{v}_i^T \lambda_j \mathbf{v}_j$$

As each of the $\mathbf{v}$'s are orthonormal, this sum is equal to $\sum_{i=1}^{d} a_i^2 \lambda_i \geq 0$.
For $\Leftarrow$: Pick $\mathbf{c} = \mathbf{v}$, some eigenvector. Then $\mathbf{v} K \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} \geq 0 \Rightarrow \lambda \geq 0$.