CSC2515: Final Exam Review

Ekansh Sharma

November 27, 2019

Built from slides by David Madras + James Lucas

Final Exam

- Time: Tuesday, Dec. 17, from 3:00pm to 6:00pm
- Location: Banting Institute, Room 131
- Office hours will be posted on the course website.

Agenda

1. A brief overview
2. Some sample questions

Basic ML Terminology

The final exam will cover everything up through Lecture 11; 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.

- Dimensionality reduction
- Clustering
- Bayes Rule
- Prior/posterior distributions
- Conjugacy
- Likelihood function
- Mahalanobis distance
- Isotropic covariance
- Conditional independence
- I.I.D.

Basic ML Terminology

The final exam will cover everything up through Lecture 11; 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 models
- Gaussian Mixture Model (GMM)
- Expectation-Maximization (EM) algorithm
- Jensen's Inequality
- Reinforcement Learning
- States/actions/rewards
- Exploration/exploitation
- Laplace Mechanism
- Sensitivity of a function
- Exponential Mechanism

Some Questions

Question 1

True of 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 of False:

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

Some Questions

Question 1

True of 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.

Some Questions

Question 1

True of 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 of 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*

Question2

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

Some Questions

Question 1

True of 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*

Question2

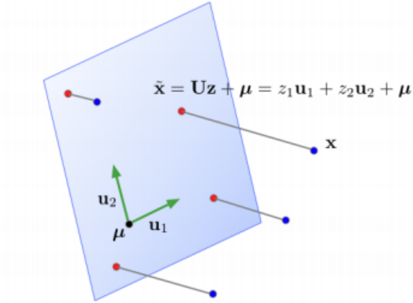
1. How can a generative model $p(x|y)$ be used as a classifier?
2. Why is dimensionality reduction useful?

Subject Areas (post midterm)

1. Principal Component Analysis
2. Probabilistic Models
3. K-Means
4. Mixture Models + EM algorithm
5. Reinforcement Learning
6. Differential Privacy

Principal Component Analysis (PCA)

1. What does PCA reconstruction minimize?
2. What is the optimal PCA subspace given empirical $\mathbf{\Sigma}$?
3. Linear and non-Linear Autoencoders


$\tilde{\mathbf{x}} = \mathbf{U}\mathbf{z} + \boldsymbol{\mu} = z_1\mathbf{u}_1 + z_2\mathbf{u}_2 + \boldsymbol{\mu}$
$\mathbf{x}$
$\mathbf{u}_2$
$\mathbf{u}_1$
$\boldsymbol{\mu}$

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

Probabilistic Models

Bayes' Rule:

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

Parameter Estimation

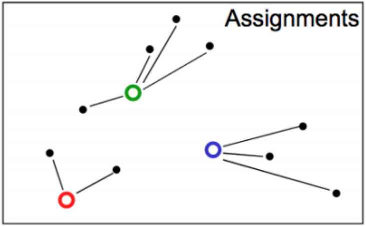
- Maximum Likelihood Estimation (MLE): $\arg\max_\theta p(\mathcal{D} \mid \theta)$
- Maximum A Posteriori Esitmation (MAP):$\arg\max_\theta p(\theta \mid \mathcal{D})$

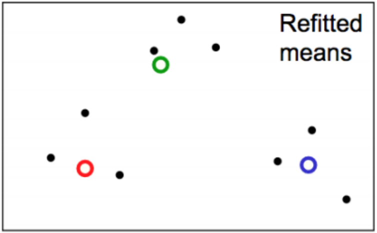
Classification

1. Generative vs Discriminative classification
2. Naive Bayes: Assumes features independent given the class
3. Gaussian Discriminant Analysis

Clustering: *K*-Means

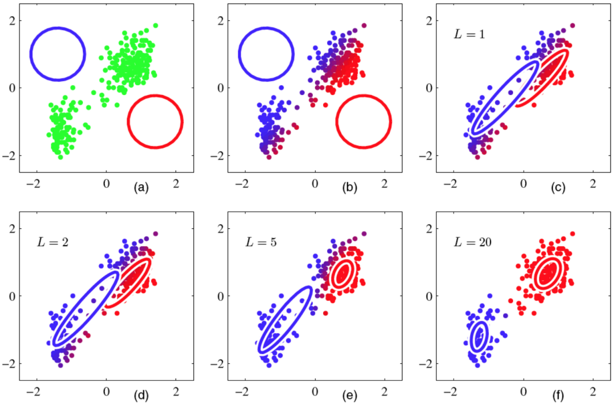
1. Initialization, assignment, refitting
2. Convergence
3. Soft vs hard K-means





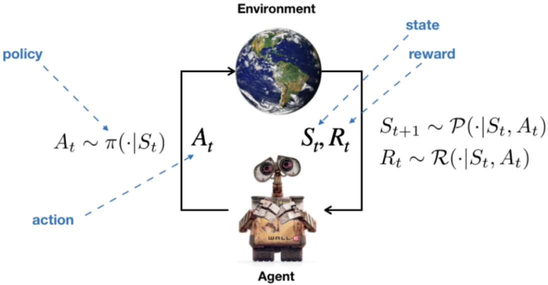
Mixture Models

1. Gaussian Mixture Model (GMM)
2. Expectation–Maximization (EM) Algorithm
   - E-Step, M-Step
   - K-means vs EM



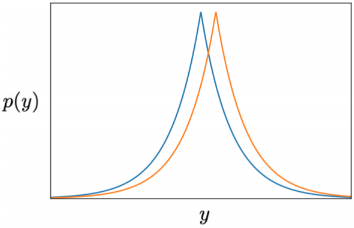
Reinforcement Learning

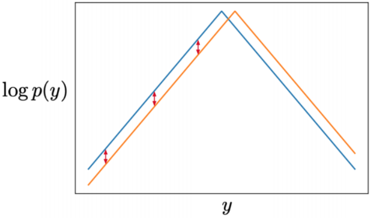
1. Choosing actions to maximize long-term reward
2. States, actions, rewards, policies, transition probability
3. Value function, Bellman Equation, value iteration
4. Q-learning
5. Exploration vs. Exploitation


Environment
policy
state
reward
$A_t \sim \pi(\cdot|S_t)$
A_t
S_t, R_t
$S_{t+1} \sim \mathcal{P}(\cdot|S_t, A_t)$
$R_t \sim \mathcal{R}(\cdot|S_t, A_t)$
action
WALL-E
Agent

Differential Privacy

1. Definition: ε-differential privacy
2. Laplace Mechanism: Add noise
3. Exponential Mechanism
4. Composition Rules





Sample Question 1

Recall that Gaussian discriminant analysis (GDA) can have very different decision boundary shapes depending on the precise model assumptions. Consider a GDA model with two classes, and where the covariance is shared between both classes and is spherical. Show mathematically that the decision boundary is linear. For reference, the multivariate Gaussian PDF is given by:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Q1 Solution

Decision boundary is when $\log p(t = 0|\mathsf{x}) = \log p(t = 1|\mathsf{x})$. That is

$$(\mathsf{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathsf{x} - \boldsymbol{\mu}_0) = (\mathsf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathsf{x} - \boldsymbol{\mu}_1) + const$$

Expanding:

$$\begin{aligned} &\mathsf{x}^T \boldsymbol{\Sigma}^{-1} \mathsf{x} - 2\mu_0^T \boldsymbol{\Sigma}^{-1} \mathsf{x} = \mathsf{x}^T \boldsymbol{\Sigma}^{-1} \mathsf{x} - 2\mu_1^T \boldsymbol{\Sigma}^{-1} \mathsf{x} + const \\ &2(\mu_1 - \mu_0)^T \boldsymbol{\Sigma}^{-1} \mathsf{x} = const \end{aligned}$$

Hence the decision boundary is linear.

Sample Question 2

We will derive the E-M update rules for a univariate Gaussian mixture model (GMM) with two mixture components. Unlike the GMMs we covered in the course, the mean μ will be shared between the two mixture components, but each component will have its own standard deviation σ_k. The model is defined as follows:

$$\begin{aligned} z &\sim \mathrm{Bernoulli}(\theta) \\ x|z = k &\sim \mathcal{N}(\mu, \sigma_k) \end{aligned}$$

(A) Write the density defined by this model (i.e. the probability of x, with z marginalized out)

(B) E-Step: Compute the posterior probability $r^{(i)} = \Pr(z^{(i)} = 1|x^{(i)})$.

(C) M-Step: Update rule for μ (keeping σ_k fixed)

(D) M-Step: Update rule for σ_1 (keeping μ fixed)

Q2 Solution

(A) $p(x) = \theta \mathcal{N}(x; \mu, \sigma_1) + (1-\theta)\mathcal{N}(x; \mu, \sigma_0)$

(B) $r^{(i)} = \frac{\theta \mathcal{N}(x^{(i)}; \mu, \sigma_1)}{\theta \mathcal{N}(x^{(i)}; \mu, \sigma_1) + (1-\theta)\mathcal{N}(x^{(i)}; \mu, \sigma_0)}$

(C)+(D) At each M-Step we optimize the following:

$$\begin{aligned}
&\mathcal{L}(\mu, \sigma_0, \sigma_1, \theta) = \\
&= \sum_{i=1}^{N} r^{(i)} \log(\mathcal{N}(x^{(i)}|\mu, \sigma_1) + r^{(i)} \log \theta \\
&+ (1 - r^{(i)}) \log \mathcal{N}(x^{(i)}|\mu, \sigma_0)) + (1 - r^{(i)}) \log(1-\theta)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} = 0 &\implies \sum_{i}^{N} r^{(i)} \frac{(x^{(i)} - \mu)}{\sigma_1^2} + (1 - r^{(i)}) \frac{(x^{(i)} - \mu)}{\sigma_0^2} = 0 \\
&\implies \sum_{i}^{N} (x^{(i)} - \mu) \left(\frac{r^{(i)}}{\sigma_1^2} + \frac{(1 - r^{(i)})}{\sigma_0^2} \right) = 0
\end{aligned}$$

Q2 Solution contd.

$$\implies \sum_{i}^{N}(x^{(i)} - \mu)\left(\sigma_0^2 r^{(i)} + \sigma_1^2(1 - r^{(i)})\right) = 0$$

Thus you get: $\mu \leftarrow \frac{\sum_{i=1}^{N} x^{(i)}\left(r^{(i)}\sigma_0^2 + (1-r^{(i)})\sigma_1^2\right)}{\sum_{i=1}^{N}\left(r^{(i)}\sigma_0^2 + (1-r^{(i)})\sigma_1^2\right)}$

$$\frac{\partial \mathcal{L}}{\partial \sigma_1^2} = 0 \implies \sigma_1^2 \leftarrow \frac{\sum_{i=1}^{N} r^{(i)}(x^{(i)} - \mu)^2}{\sum_{i=1}^{N} r^{(i)}}$$