EECS 545: Machine Learning

Lecture 23. Midterm review

Honglak Lee and Michał Dereziński

04/04/2022

Midterm logistics, syllabus and final grade

- Thursday 4/7 at 6pm-8pm EST.
- Room assignments (based on last name initials):
  - Last name initials, Room
  - A-M, CHRYS 220
  - N-X, DOW 1013
  - Y-Z, EECS 1200
- Grading scheme = **Maximum** of Scheme 1 and Scheme 2
  - Scheme 1: HW 30%, Midterm 30%, Project 40%
  - Scheme 2: HW 35%, Midterm 20%, Project 45%
- Syllabus for midterm: Lecture 1-16 + Lecture 19 (ML Advice)

Examples

- Which of the following are **linear classifiers**? Choose all options that are correct.
  a) Logistic regression
  b) SVM with kernel $k(x, y) = x^T y$
  c) Gaussian discriminant analysis
  d) SVM with RBF kernel.

Examples

- Which of the following are **linear classifiers**? Choose all options that are correct.
  a) **Logistic regression**
  b) **SVM with kernel k(x, y) = $x^T y$**
  c) **Gaussian discriminant analysis**
  d) SVM with RBF kernel.

Examples

- Which of the following are **max-margin classifiers?** Choose all options that are correct.
  a) Logistic regression
  b) SVM with kernel $k(x, y) = x^T y$
  c) Gaussian discriminant analysis
  d) SVM with RBF kernel.

Examples

- Which of the following are **max-margin classifiers**? Choose all options that are correct.
  a) Logistic regression
  b) SVM with kernel $k(x, y) = x^T y$
  c) Gaussian discriminant analysis
  d) SVM with RBF kernel.

Examples

- Write the feature map $\varphi(x)$ associated with the homogenous quadratic kernel $k(x, y) = (x^T y)^2$ where $x = [x_1\ x_2]^T \in R^2$ and $y = [y_1\ y_2]^T \in R^2$.

Examples:

- Write the feature map $\varphi(x)$ associated with the homogenous quadratic kernel $k(x, y) = (x^T y)^2$ where $x = [x_1\ x_2]^T \in R^2$ and $y = [y_1\ y_2]^T \in R^2$.

- $\boldsymbol{\varphi(x) = [x_1^2 \quad x_2^2 \qquad \sqrt{2}\, x_1 x_2]^T}$

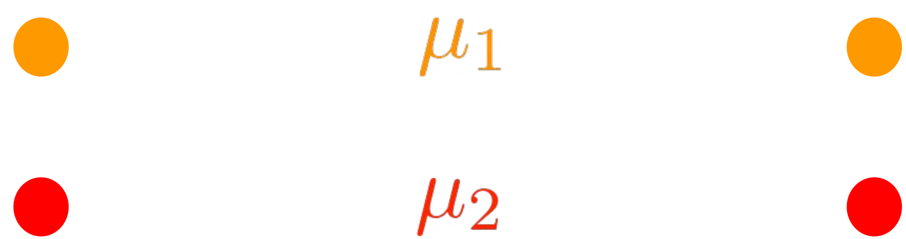
Examples

- True/False: k-means algorithm always converges to the global optimum solution.

Examples

- True/False: k-means algorithm always converges to the global optimum solution.

- **False. Can get stuck in local minimum.**



Examples

- Which of the following best describes what discriminative approaches try to model? (w are the parameters in the model)
  a) $p(y|x, w)$
  b) $p(y, x)$
  c) $p(x|y, w)$
  d) $p(w|x, y)$

Source: https://www.seas.upenn.edu/~cis520/exams/midterm_2016_solns.pdf

Examples

- Which of the following best describes what discriminative approaches try to model? (w are the parameters in the model)
  a) **$p(y|x, w)$**
  b) $p(y, x)$
  c) $p(x|y, w)$
  d) $p(w|x, y)$

Source: https://www.seas.upenn.edu/~cis520/exams/midterm_2016_solns.pdf

Examples

A random variable follows an exponential distribution with parameter λ ($\lambda > 0$) if it has the following density:

$$p(t) = \lambda e^{-\lambda t},\ t \in [0, \infty)$$

Imagine you are given i.i.d. data $D = (t_1, \ldots, t_n)$ where each t_i is drawn from an exponential distribution with parameter λ.

- Compute the log-likelihood of data: $\log p(D|\lambda)$
- Compute the Maximum Likelihood (ML) estimate of λ.

Source: https://courses.cs.washington.edu/courses/cse546/14au/exams/14au_midterm_sol.pdf

Examples

- **Compute the log-likelihood of data: log p(D|λ)**

$$\begin{aligned}\ln p(T) &= \ln \prod_i p(t_i) \\ &= \sum_i \ln(\lambda e^{-\lambda t_i}) \\ &= \sum_i \ln \lambda - \lambda t_i \\ &= \boxed{n \ln \lambda - \lambda \sum_i t_i}\end{aligned}$$

Source: https://courses.cs.washington.edu/courses/cse546/14au/exams/14au_midterm_sol.pdf

Examples

- **Compute the ML estimate of λ.**

$$\frac{\partial}{\partial \lambda}(n \ln \lambda - \lambda \sum_i t_i) = 0$$

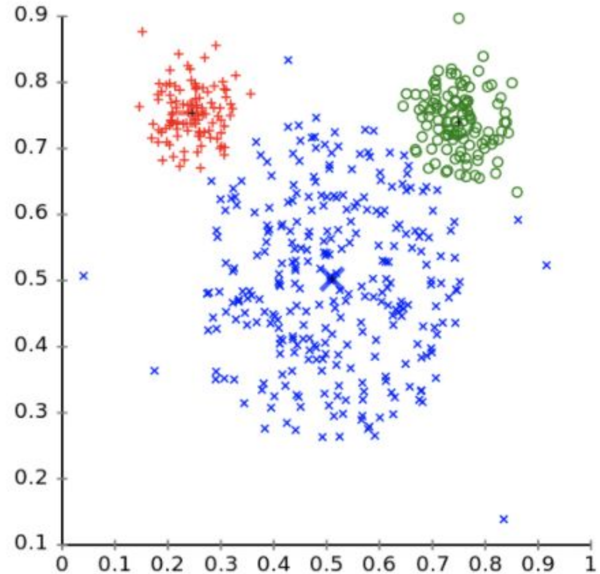
$$\frac{n}{\lambda} - \sum_i t_i = 0$$

$$\hat{\lambda}_{MLE} = \boxed{\frac{n}{\sum_i t_i}}$$

Source: https://courses.cs.washington.edu/courses/cse546/14au/exams/14au_midterm_sol.pdf

Examples

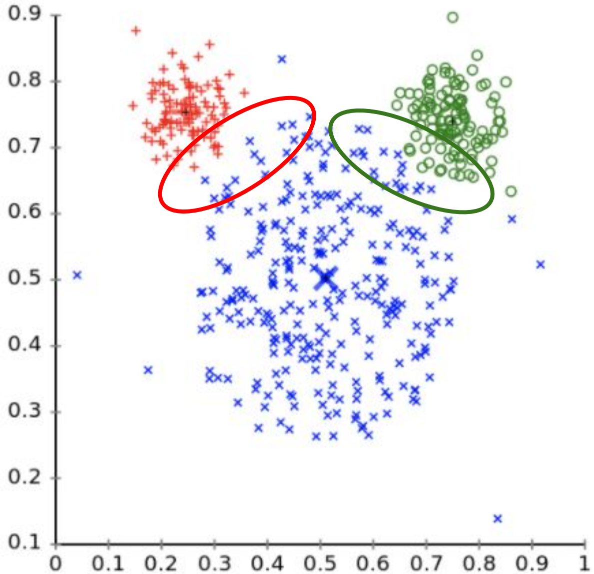
- True/False: The following clustering is achievable with K-means algorithm (K=3).



Source: https://huyenchip.com/ml-interviews-book/contents/8.1.2-questions.html

Examples:

- True/False: The following clustering is achievable with K-means algorithm (K=3).



False, assignment is done wrt distance from means, so the samples in the ovals will be of different color. Basically, can’t have different covariance matrix. Possible with GMM though.

Examples

- True/False: Logistic loss is better than L2 loss in classification tasks.

Source: http://alex.smola.org/teaching/10-701-15/exam/final.pdf

Examples

- True/False: Logistic loss is better than L2 loss in classification tasks.

- **Answer: True. Correctly classified points that are far away from the decision boundary have much less impact on the decision boundary**

Source: http://alex.smola.org/teaching/10-701-15/exam/final.pdf

Examples

- True/False: A Multilayer perceptron (MLP) with activation function $f(x) = 0.6x + 2$ is a good choice for classification problems.

Examples

- True/False: A Multilayer perceptron (MLP) with activation function $f(x) = 0.6x + 2$ is a good choice for classification problems.
- **False. The MLP reduces to a linear classifier.**

Examples:

- Which of the following are true for RNNs? Choose all options that are correct
  a) RNNs can be used for time series forecasting.
  b) RNNs can be easily parallelized.
  c) RNNs can be used for language translation.
  d) RNNs can suffer from vanishing/exploding gradients.

- True/False: It is easier to avoid exploding gradients than vanishing gradients.

Examples:

- Which of the following are true for RNNs? Choose all options that are correct
  a) **RNNs can be used for time series forecasting.**
  b) RNNs can be easily parallelized.
  c) **RNNs can be used for language translation.**
  d) **RNNs can suffer from vanishing/exploding gradients.**

- True/False: It is easier to avoid exploding gradients than vanishing gradients.

  True. You can just clip large gradients.

Examples:

- True/False: Forward propagation is required during training as well as testing, while Backward propagation is done only during training.

- True/False: The forward propagation of BatchNormalization layer involves the same computation steps during training as well as testing.

Examples:

- True/False: Forward propagation is required during training as well as testing, while Backward propagation is done only during training.
- **True**.

- True/False: The forward propagation of BatchNormalization layer involves the same computation steps during training as well as testing.
- **False. During training, we compute batch statistics for mu and sigma. But during testing, we use precomputed global statistics.**

Examples:

- Given a input of size (1, 3, 224, 224), calculate the output size at each intermediate layer of a CNN given below:

Layer	Output size
Conv(outChannels=64, inChannels=3, kernel=(3,3), stride=2, padding=1)	?
MaxPool2D(kernel_size=(2,2), stride=2)	?
Conv(outChannels=1, inChannels=64, kernel=(5, 5), stride=1, padding=0)	?

Examples:

- Given a input of size (1, 3, 224, 224), calculate the output size at each intermediate layer of a CNN given below:

Layer	Output size
Conv(outChannels=64, inChannels=3, kernel=(3,3), stride=2, padding=1)	(1, 64, 112, 112)
MaxPool2D(kernel_size=(2,2), stride=2)	(1, 64, 56, 56)
Conv(outChannels=1, inChannels=64, kernel=(5, 5), stride=1, padding=0)	(1, 1, 52, 52)

$$\text{Output\_dim} = \left[(\text{input\_dim} - \text{kernel} + 2 * \text{padding})/\text{stride} + 1\right]$$

(Floor function)

Examples:

- Which of the following statements are true about the EM algorithm?
  a) EM can be used to estimate the parameters of a latent variable generative model.
  b) The log-likelihood increases monotonically during the course of EM algorithm
  c) The EM algorithm converges to the global optimum solution.

Examples:

- Which of the following statements are true about the EM algorithm?
  a) **EM can be used to estimate the parameters of a latent variable generative model.**
  b) **The lower bound of log-likelihood increases monotonically during the course of EM algorithm**
  c) The EM algorithm converges to the global optimum solution.

Examples:

In this problem, you will derive an EM algorithm for estimating the mixing parameter for a mixture of arbitrary probability densities f_1 and f_2. For example, $f_1(x)$ could be a standard normal distribution centered at 0, and $f_2(x)$ could be the uniform distribution between $[0, 1]$. You can think about such mixtures in the following way: First, you flip a coin. With probability λ (i.e., the coin comes up heads), you will sample x from density f_1, with probability $(1 - \lambda)$ you sample from density f_2.

More formally, let $f_\lambda(x) = \lambda f_1(x) + (1-\lambda) f_2(x)$, where f_1 and f_2 are arbitrary probability density functions on $\mathbb{R}$, and $\lambda \in [0, 1]$ is an unknown mixture parameter.

1. Given a data point x, and a value for the mixture parameter λ, compute the probability that x was generated from density f_1.

- p(c=1|x) = p(x|c=1) p(c = 1) / (p(x|c=1)p(c=1) + p(x|c=0)p(c=0))

$$\boxed{\frac{\lambda f_1(x)}{\lambda f_1(x) + (1-\lambda) f_2(x)}}$$

Source: https://www.cs.cmu.edu/~guestrin/Class/10701-F07/additional/final-s2006-sol.pdf

Examples

In this problem, you will derive an EM algorithm for estimating the mixing parameter for a mixture of arbitrary probability densities f_1 and f_2. For example, $f_1(x)$ could be a standard normal distribution centered at 0, and $f_2(x)$ could be the uniform distribution between $[0, 1]$. You can think about such mixtures in the following way: First, you flip a coin. With probability λ (i.e., the coin comes up heads), you will sample x from density f_1, with probability $(1 - \lambda)$ you sample from density f_2.

More formally, let $f_\lambda(x) = \lambda f_1(x) + (1-\lambda) f_2(x)$, where f_1 and f_2 are arbitrary probability density functions on $\mathbb{R}$, and $\lambda \in [0, 1]$ is an unknown mixture parameter.

2. Now you are given a data set $\{x_1, \ldots, x_n\}$ drawn i.i.d. from the mixture density, and a set of coin flips $\{c_1, c_2, \ldots, c_n\}$, such that $c_i = 1$ means that x_i is a sample from f_1, and $c_i = 0$ means that x_i was generated from density f_2. For a fixed parameter λ, compute the complete log-likelihood of the data, i.e., $\log P(x_1, c_1, x_2, c_2, \ldots, x_n, c_n \mid \lambda)$.

$$
\begin{aligned}
\log P(x_1, c_1, \ldots, x_n, c_n|\lambda) &= \sum_{i=1}^{N} \log P(x_i, c_i|\lambda) \\
&= \sum_{i=1}^{N} \log \left[(\lambda f_1(x))^{c_i} \quad ((1-\lambda) f_2(x))^{1-c_i}\right] \\
&= \sum_{i=1}^{N} c_i \left[\log \lambda + \log f_1(x)\right] + (1 - c_i)\left[\log(1-\lambda) + \log f_2(x)\right]
\end{aligned}
$$

Examples:

In this problem, you will derive an EM algorithm for estimating the mixing parameter for a mixture of arbitrary probability densities f_1 and f_2. For example, $f_1(x)$ could be a standard normal distribution centered at 0, and $f_2(x)$ could be the uniform distribution between $[0, 1]$. You can think about such mixtures in the following way: First, you flip a coin. With probability λ (i.e., the coin comes up heads), you will sample x from density f_1, with probability $(1 - \lambda)$ you sample from density f_2.

More formally, let $f_\lambda(x) = \lambda f_1(x) + (1-\lambda) f_2(x)$, where f_1 and f_2 are arbitrary probability density functions on $\mathbb{R}$, and $\lambda \in [0, 1]$ is an unknown mixture parameter.

3. Now you are given only a sample $\{x_1, \dots, x_n\}$ drawn i.i.d. from the mixture density, without the knowledge about which component the samples were drawn from (i.e., the c_i are unknown). Using your derivations from part 1 and 2, derive the E- and M-steps for an EM-algorithm to compute the Maximum Likelihood Estimate of the mixture parameter λ. Please describe your derivation of the E- and M-step clearly in your answer.

Examples:

E - step: Calculate the following for all i

$$\gamma_i = p(c=1|x_i) = \frac{\lambda f_1(x)}{\lambda f_1(x) + (1-\lambda) f_2(x)}$$

M - step:

$$\arg\max_{\lambda} \sum_{i=1}^{N} \gamma_i \log P(x_i, c_i = 1) + (1-\gamma_i) \log P(x_i, c_i = 0)$$

$$\arg\max_{\lambda} \sum_{i=1}^{N} \gamma_i \left[\log \lambda + \log f_1(x)\right] + (1-\gamma_i) \left[\log(1-\lambda) + \log f_2(x)\right]$$

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} \gamma_i$$

Examples from practice exam:

- (Q2) Assume we have the following optimization problem

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 + C_+ \sum_{i:t^{(i)}=1} \xi^{(i)} + C_- \sum_{i:t^{(i)}=-1} \xi^{(i)} \\ \text{subject to} \quad & t^{(i)}(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi^{(i)},\ i = 1, \ldots, N \\ & \xi^{(i)} \geq 0,\ i = 1, \ldots, N \end{aligned}$$

Eliminate the slack variable to get a simplified objective function E(w, b). Specifically, fill in the blanks given below to complete the expression for E(w, b).

$$\frac{1}{2}||\mathbf{w}||^2 + C_+ \boxed{} + C_- \boxed{}$$

Examples from practice exam:

- (Q2) Assume we have the following optimization problem

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 + C_+ \sum_{i:t^{(i)}=1} \xi^{(i)} + C_- \sum_{i:t^{(i)}=-1} \xi^{(i)} \\ \text{subject to} \quad & t^{(i)}(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi^{(i)}, \ i = 1, \ldots, N \\ & \xi^{(i)} \geq 0, \ i = 1, \ldots, N \end{aligned}$$

Eliminate the slack variable to get a simplified objective function E(w, b). Specifically, fill in the blanks given below to complete the expression for E(w, b).

$$\boxed{\frac{1}{2}\|\mathbf{w}\|^2 + C_+ \sum_{i:t^{(i)}=1} \max\left\{0, 1 - \left(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b\right)\right\} + C_- \sum_{i:t^{(i)}=-1} \max\left\{0, 1 + \left(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b\right)\right\}}$$

Examples from practice exam:

- (Q2 continued) Complete the expression for the gradient $\nabla_{w} E$

$$\mathbf{w} - C_{+} \sum_{i:t^{(i)}=1} \boxed{} \phi(\mathbf{x}^{(i)}) + C_{-} \sum_{i:t^{(i)}=-1} \boxed{} \phi(\mathbf{x}^{(i)})$$

$$\begin{aligned}
\nabla_{\mathbf{w}} E(\mathbf{w}, b) & \\
= \; & \mathbf{w} - C_{+} \sum_{i:t^{(i)}=1} \mathbf{I}\left[(\mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b) \geq 1\right] \mathbf{0} - C_{+} \sum_{i:t^{(i)}=1} \mathbf{I}\left[(\mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b) < 1\right] \phi(\mathbf{x}^{(i)}) \\
& + C_{-} \sum_{i:t^{(i)}=-1} \mathbf{I}\left[(\mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b) \leq -1\right] \mathbf{0} + C_{-} \sum_{i:t^{(i)}=-1} \mathbf{I}\left[(\mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b) > -1\right] \phi(\mathbf{x}^{(i)}) \\
= \; & \mathbf{w} - C_{+} \sum_{i:t^{(i)}=1} \mathbf{I}\left[(\mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b) < 1\right] \phi(\mathbf{x}^{(i)}) + C_{-} \sum_{i:t^{(i)}=-1} \mathbf{I}\left[(\mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b) > -1\right] \phi(\mathbf{x}^{(i)})
\end{aligned}$$

Examples from practice exam:

- (Q1) Complete the proof for 1a.

Consider the training data $\{(\mathbf{x}^{(1)}, t^{(1)}), \cdots, (\mathbf{x}^{(N)}, t^{(N)})\}$ where $\mathbf{x} \in \mathbb{R}^D$ and $t \in \mathbb{R}$. Assume that the output t is generated from input $\mathbf{x}$ as follows:

$$\begin{aligned} t &= \mathbf{w}^T \phi(\mathbf{x}) + \epsilon \\ \epsilon &\sim \text{Laplace}(\epsilon; 0, 1) \end{aligned}$$

where the probability density function of Laplace distribution is given as

$$\text{Laplace}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

(a) [4 points] Show that the Maximum Likelihood Estimation (MLE) of $\mathbf{w}$ for the data (i.e., maximizing the log-likelihood of t conditioned on $\mathbf{x}$ over the training data) is equivalent to the "robust linear regression" problem, which is written as

$$\min_{\mathbf{w}} \sum_{i=1}^{N} |t^{(i)} - \mathbf{w}^T \phi(\mathbf{x}^{(i)})|$$

Examples from practice exam:

- (Q1) Complete the proof for 1a.

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \log P\left(t^{(i)} | \mathbf{x}^{(i)}\right) \\
&= \boxed{} \\
&= \boxed{} \\
&= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left| t^{(i)} - \mathbf{w}^T \phi\left(\mathbf{x}^{(i)}\right) \right|
\end{aligned}
$$

Examples from practice exam:

- (Q1) Complete the proof for 1a.

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \log P\left(t^{(i)} | \mathbf{x}^{(i)}\right) \\
&= \arg\max_{\mathbf{w}} \quad \frac{1}{N} \sum_{i=1}^{N} \log \frac{1}{2} \exp\left(-\left|t^{(i)} - \mathbf{w}^T \phi\left(\mathbf{x}^{(i)}\right)\right|\right) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} -\left|t^{(i)} - \mathbf{w}^T \phi\left(\mathbf{x}^{(i)}\right)\right| \\
&= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left|t^{(i)} - \mathbf{w}^T \phi\left(\mathbf{x}^{(i)}\right)\right|
\end{aligned}
$$

Examples

- Given the following constrained optimization problem. Write the expression for:
  - Lagrangian function L(ξ, w, λ)
  - Lagrangian dual function $\tilde{L}(\lambda)$

$$\min_{\xi, w} \xi^3 + \frac{1}{2} w^2 \quad \text{s.t.}\; aw + b \leq \xi$$

Examples

- Given the following constrained optimization problem. Write the expression for:
  - **Lagrangian function L(ξ, w, λ).**

$$\min_{\xi,w} \xi^3 + \frac{1}{2}w^2 \quad \text{s.t. } aw + b \leq \xi$$

$$L(\xi, w, \lambda) = \xi^3 + \frac{1}{2}w^2 + \lambda(aw + b - \xi)$$

Examples

- Given the following constrained optimization problem. Write the expression for:
  - Lagrangian function L(ξ, w, λ).
  - **Lagrangian dual function** $\tilde{L}(\lambda)$

$$\min_{\xi,w} \xi^3 + \frac{1}{2}w^2 \quad \text{s.t. } aw + b \leq \xi \qquad L(\xi, w, \lambda) = \xi^3 + \frac{1}{2}w^2 + \lambda(aw + b - \xi)$$

$$\tilde{L}(\lambda) = \min_{\xi,w} L(\xi, w, \lambda)$$

$$\frac{\partial L(\xi, w, \lambda)}{\partial w} = 0 \implies w = -\lambda a \qquad \frac{\partial L(\xi, w, \lambda)}{\partial \xi} = 0 \implies \xi = \left(\frac{\lambda}{3}\right)^{1/2}$$

$$\tilde{L}(\lambda) = \lambda^{3/2}\left(1 - \frac{1}{\sqrt{3}}\right) - \frac{1}{2}\lambda^2 a^2 + \lambda b$$