Midterm Review Tutorial

Mohammadreza Safavi
mohammadreza.safavi@mail.utoronto.ca

University of Toronto
ECE1513: Introduction to Machine Learning
Instructor: Dr. Ali Bereyhi

Feb 24, 2025

Design Problem: Predicting Heart Disease

Question 1: You want to design a data-driven solution for this problem: given the results of a blood test, what is the chance that the patient is developing a specific heart condition.

- Formulate this problem as a supervised learning problem.
- Explain how you could make the three components: Dataset, Model, Learning Algorithm.
- Explain how you can train this model. You could specify the loss function if you need.
- Explain how you can use the trained model to predict the chance of heart condition for a new patient.

Formulating the Problem

Answer:

- **Problem Formulation:** This is a supervised learning classification task where the goal is to predict the probability of a heart condition based on blood test results.

Components of the Solution

Answer:

- **Dataset:** Blood test data with labels indicating the presence or absence of the heart condition.
- **Model:** Choose a model such as logistic regression, decision trees, or neural networks.
- **Learning Algorithm:** Use optimization algorithms like stochastic gradient descent (SGD) with a suitable loss function, such as cross-entropy loss.

Training the Model

Answer:

- **Training:** Define the loss function (e.g., cross-entropy loss), optimize the model parameters using the training data, and validate the model using a separate validation set to ensure it generalizes well.

Using the Trained Model

Answer:

- **Prediction:** Use the trained model to predict the probability of a heart condition for new patient data by inputting their blood test results into the model.

K-Means Clustering: Initial Assignment

Question 2: We have the following points in our dataset $D = \{1, 2, 3, 10, 17, 20\}$. We start K-means clustering with $K = 2$ clusters. The initial centroids are $\mu_1 = 0$ and $\mu_2 = 5$.

- **Part 1: Cluster Assignments:** Assign each point to the nearest centroid.
    - Points assigned to $\mu_1 = 0$: $\{1, 2\}$
    - Points assigned to $\mu_2 = 5$: $\{3, 10, 17, 20\}$

K-Means Clustering: First Iteration

Part 2: Compute the updated centroids after the first iteration.

- **Updated Centroids:**
  - $\mu_1 = \frac{1+2}{2} = 1.5$
  - $\mu_2 = \frac{3+10+17+20}{4} = 12.5$

K-Means Clustering: Converging Centroids

Part 3: Specify the converging centroids.

- **Iteration 2:**
  - Points assigned to $\mu_1 = 1.5$: $\{1, 2, 3\}$
  - Points assigned to $\mu_2 = 12.5$: $\{10, 17, 20\}$
  - Updated centroids:
    - $\mu_1 = \frac{1+2+3}{3} = 2$
    - $\mu_2 = \frac{10+17+20}{3} = 15.67$
- **Iteration 3:**
  - Points assigned to $\mu_1 = 2$: $\{1, 2, 3\}$
  - Points assigned to $\mu_2 = 15.67$: $\{10, 17, 20\}$
  - Updated centroids:
    - $\mu_1 = \frac{1+2+3}{3} = 2$
    - $\mu_2 = \frac{10+17+20}{3} = 15.67$
- **Converged Centroids:** $\mu_1 = 2$, $\mu_2 = 15.67$

Likelihood of λ for Exponential Distribution

Question 3: Determine the likelihood of λ for the given dataset $D = \{1, 2, 3\}$ assuming an exponential distribution with PDF:

$$P_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Answer: The likelihood function for λ given the dataset is:

$$L(\lambda; D) = \prod_{i=1}^{3} \lambda e^{-\lambda x_i} = \lambda^3 e^{-\lambda \sum_{i=1}^{3} x_i}$$

Log-Likelihood Function

Part 2: Specify the log-likelihood function using the natural logarithm (ln).
Answer: The log-likelihood function is:

$$\ln L(\lambda; D) = \ln(\lambda^3 e^{-\lambda \sum_{i=1}^{3} x_i}) = 3 \ln \lambda - \lambda \sum_{i=1}^{3} x_i$$

Maximum Likelihood Estimate of λ

Part 3: Find the maximum likelihood estimate of λ.
Answer: To find the MLE, take the derivative of the log-likelihood and set it to zero:

$$\frac{d}{d\lambda}\left(3\ln\lambda - \lambda\sum_{i=1}^{3} x_i\right) = \frac{3}{\lambda} - \sum_{i=1}^{3} x_i = 0$$

Solving for λ, we get:

$$\lambda^* = \frac{3}{\sum_{i=1}^{3} x_i} = \frac{3}{6} = 0.5$$

PCA: Algorithm Formulation

Question 4: Formulate the PCA as an algorithm, i.e., given a dataset D, explain how PCA performs dimensionality reduction. Practice through the following items:

- Write a simple pseudo-code for PCA whose input is the dataset D and the latent dimension K.

Answer:

1. Form the sample covariance matrix Σ.
2. Decompose Σ to find eigenvalues and eigenvectors.
3. Select the top K principal eigenvectors.
4. Form matrix U and vector μ.

PCA: Pseudo-code

Pseudo-code for PCA:

- **Input:** Dataset $D \in \mathbb{R}^{n \times d}$, latent dimension K
- **Output:** Reduced dataset $D' \in \mathbb{R}^{n \times K}$

Algorithm 1 PCA Algorithm

```
Input: Dataset $D \in \mathbb{R}^{n \times d}$, latent dimension $K$
Output: Reduced dataset $D' \in \mathbb{R}^{n \times K}$
Compute the mean of the dataset: $\mu = \frac{1}{n} \sum_{i=1}^{n} D_i$
Center the dataset: $\tilde{D} = D - \mu$
Compute the covariance matrix: $\Sigma = \frac{1}{n} \tilde{D}^T \tilde{D}$
Perform eigen decomposition on $\Sigma$: $\Sigma = V \Lambda V^T$
Select the top $K$ eigenvectors to form matrix $U \in \mathbb{R}^{d \times K}$
Transform the dataset: $D' = \tilde{D} U$
Return: $D'$
```

PCA: Reconstruction Error

Question: When is the reconstruction error absolutely zero?
Answer: The reconstruction error is absolutely zero when we have at most K non-zero eigenvalues for the sample covariance matrix.

Linear Regression: Empirical Risk Minimization

Question 5: Formulate the empirical risk minimization using squared loss function $L(y, v) = (y - v)^2$.

Answer: The empirical risk for linear regression can be formulated as:

$$R(w) = \sum_{i=1}^{n} (w^T x_i - v_i)^2$$

Linear Regression: Empirical Risk as Vector Norm

Part 2: Represent the empirical risk as a vector norm, i.e., $\|X^T w - v\|^2$.
Answer: Putting the data points in matrix form X and the labels in vector form y, we have $X^T w - y$ as the residual vector. Hence the empirical risk can be represented as: The empirical risk can be represented as:

$$R(w) = \|X^T w - v\|^2$$

Linear Regression: Gradient and Optimal Weight Vector

Part 3: Use the fact $\nabla \|X^T w - v\|^2 = 2XX^T w - 2Xv$ to find the optimal choice of w.
Answer: Setting the gradient to zero to minimize the risk:

$$\nabla R(w) = 2XX^T w - 2Xv = 0$$

Solving for w, we get:

$$w^* = (XX^T)^{-1} Xv$$

Linear Regression: Validity of the Solution

Part 4: Discuss when this solution is valid.
Answer: The solution is valid when XX^T is invertible, i.e., all its eigenvalues are positive (no zero eigenvalue).

Gradient Descent: Update Rule

Question 6: State the update rule of gradient descent algorithm and specify its components.

Answer:

- Update rule: $w \leftarrow w - \eta \nabla J(w)$
- Components:
  - w: Current weight vector
  - η: Learning rate
  - $\nabla J(w)$: Gradient of the cost function

Gradient Descent: Convergence Point

Part 2: What is the property of the converging point of gradient descent algorithm?
Answer: It's a stationary point.

Gradient Descent: Convergence to Minimum

Part 3: Explain why gradient descent always converges to a minimum point.

Answer: We can see that close to any stationary point other than the minimum, the algorithm pushes us outwards (under certain conditions).

Gradient Descent: Learning Rate

Part 4: What are the pros and cons of large learning rate?

Answer:

- Pros:
  - Speeds up convergence.
- Cons:
  - May cause instability and overshooting.

Gradient Descent: Logistic Regression

Part 5: Does using gradient descent for logistic regression send us towards optimal solution? Explain your reason.
Answer: Yes, the risk function is convex in this case.

Interpreting the Components of SVC Optimization

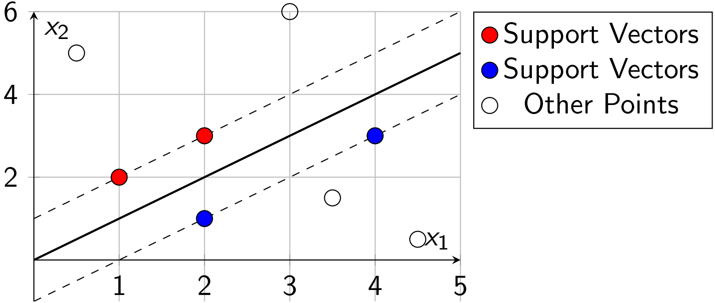
Question 7: Try to have a clear understanding of the support vector classifier. To this end, recall the SVC training problem which is:

$$\min \|w\|^2 \quad \text{subject to} \quad v_n w^T x_n \geq 1 \quad \text{for all } n$$

- **Purpose of minimizing the objective:** To maximize the margin, since the margin is proportional to $\frac{1}{\|w\|^2}$.
- **Purpose of the constraints:** To guarantee zero classification error on the dataset.

Support Vectors and Maximum Margin

- **Defining support vectors using the solution:** For support vectors, the constraint holds with equality.
- **Intuitive definition of support vectors:** They are the closest vectors to the boundary.
- **Importance of maximum margin:** We want to improve generalization (confidence).



SVM Numerical Example: XOR Dataset

Question 8: Try a sample numerical example, where by going to higher dimensions the dataset gets linearly separable, like those in Lecture 6.

Answer:

- Consider the XOR dataset: $\{(0,0),(0,1),(1,0),(1,1)\}$ with labels $\{-1,1,1,-1\}$.
- This dataset is not linearly separable in 2D.

Transforming XOR Dataset

Answer (continued):

- Apply the transformation $\phi(x_1, x_2) = (x_1, x_2, x_1 x_2)$.
- Transformed dataset: $\{(0,0,0),(0,1,0),(1,0,0),(1,1,1)\}$.
- In this 3D space, the dataset is linearly separable.

Linear Separation in Higher Dimension

Answer (continued):

- The separating hyperplane can be defined as $x_1 + x_2 - 2x_3 - 0.5 = 0$.
- Points $(0, 0, 0)$ and $(1, 1, 1)$ lie on one side of the hyperplane.
- Points $(0, 1, 0)$ and $(1, 0, 0)$ lie on the other side.

Good luck with your midterm!