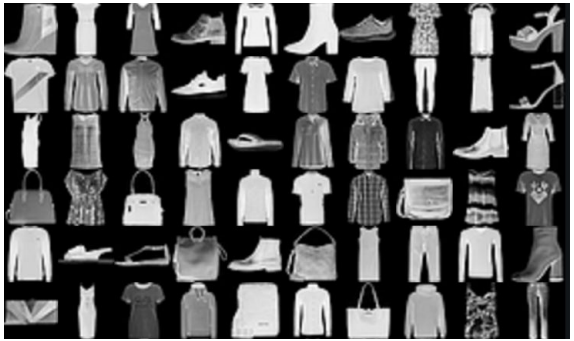
CSE 575: Statistical Machine Learning (Spring 2021)
Instructor: Nupur Thakur

Supervised Learning

Table of contents

Project Part 1 - Density Estimation and Supervised Learning

- **Due on: 17 Feb, 11:59 PM MST.**
- Binary classification using Fashion-MNIST dataset.
- Two classes - T-shirt (label 0) and Trouser (label 1)
- 28x28 grayscale images
- Training set: "Tshirt": 6000; "Trouser": 6000.
- Testing set: "Tshirt": 6000; "Trouser": 1000.



More information on Fashion-MNIST, refer https://github.com/zalandoresearch/fashion-mnist

Project Part 1 - Tasks

- Calculate two features of each image - average and standard deviation.
- Estimate parameters for 2D Gaussian distribution for each class.
- Implement Naive Bayes to perform classification
- Train Logistic regression using gradient ascent to perform classification

Project Part 1 - Deliverables

- Code:
  - Acceptable file types are .py/.m or .zip.
  - Well-commented code. Be sure to read through the directions carefully to ensure you have included all necessary parts in your code.
- Report:
  - Acceptable File types: .pdf
  - Length: 2-5 A4 pages
  - Content: Refer Canvas

Project Part 1 - Languages/Software

- What can you use?
  - Python/MATLAB to code.
  - Libraries or in-built functions to manipulate the data i.e. NumPy, SciPy libraries etc.
- What cannot be used?
  - Packages like scikit-learn which have ready-to-use algorithms.

What is supervised learning?

- Data - <sample, label> pairs
- Objective - Learn from the given data such that label can be predicted for a new data sample.
- Based on labels - → Regression
  → Classification

Density Estimation

- Estimating underlying probability density function, based on the training data
- Parametric: each class of images (the feature vectors) may be modeled by a density function $p_\theta(x)$ with parameter θ.
- Non-parametric: makes no assumption about the distribution of data or modeling the data without any parameters

Maximum Likelihood Estimation

- Given some training data and assuming a parametric model p(x|θ); what specific θ will fit/explain the data best?
- To consider all the samples denoted by D={x_1, x_2, ……, x_n}, assume that all the samples are i.i.d - independent and identically distributed.
- So, data likelihood represented by L(θ) is -

$$L(\boldsymbol{\theta}) = P(D|\boldsymbol{\theta}) = \prod_i P(x_i|\theta)$$

Maximum Likelihood Estimation

1. Likelihood function
2. Take derivative
3. Set to 0.

- Maximum Likelihood Estimation (MLE): Finding the parameter that maximizes the likelihood function

$$\hat{\theta} = \text{argmax}_{\theta} p(\mathbf{x}|\theta)$$

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i}^{n} P(x_i|\theta)$$

$$= \underset{\theta}{\text{argmax}} \log \prod_{i}^{n} P(x_i|\theta) = \underset{\theta}{\text{argmax}} \sum_{i} \log P(x_i|\theta)$$

MLE - Example 1

$\theta = \{\mu, \sigma^2\}$

Given *n* i.i.d. samples $\{x_i\}$ from the 1-D normal distribution $N(\mu, \sigma^2)$, find the MLE for μ and σ^2.

1. Likelihood func

$$p(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

$$= \prod_{i=1}^{n}$$

$$\log p(D|\theta) = \log \prod_{i=1}^{n} p(x_i|\theta) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$-n \log \sigma\sqrt{2\pi}$$

MLE - Example 1

(μ, σ^2)

(2) Take derivative w.r.t μ

$$\frac{\partial}{\partial \mu} \log P(D|\theta) = 0 - \sum_{i=1}^{n} \frac{-2(x_i - \mu)}{2\sigma^2}$$

$$= \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2}$$

(3) Equate derivative to 0.

$$\sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$n\mu = \sum_{i=1}^{n} x_i$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \text{Sample mean}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Generative vs Discriminative Models

- Generative models -
  - Learn P(y) and P(x|y).
  - Ex: Bayesian classifier, Naive Bayes.
- Discriminative models -
  - Directly learn P(y|x)
  - Ex: Logistic Regression

Questions?