Weekly Report 9

The week: Oct 28, Oct 29

Utkarsh Patel
18EC30048

Topics Covered

1 Parametric Methods

- Probability density functions of known distributions and a set of parameters are used and estimated.
- Bayesian inference is then applied to devise conclusions.

2 Maximum Likelihood Estimation

- Let θ be any parameter and $X = \{x^t\}$ be the training samples. Then, likelihood of parameter given training examples is defined as
$$l(\theta|X) = P(X|\theta) = \prod_t P(x^t|\theta)$$
The MLE estimate of θ is given as
$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, l(\theta|X)$$
- It is better to define log-likelihood and this eases out the computation.

3 ML Estimation under Bernoulli distribution

- Two possible outcomes are possible for each instance. Let those two outcomes be 0 and 1. Let the probability of 1 be p. Then probability function can be written as
$$P(x) = p^x(1-p)^{1-x}$$
Then, $E(x) = p, Var(x) = p(1-p)$.
The log-likelihood would be
$$L(p|X) = log \prod_t p^{x^t}(1-p)^{1-x^t}$$
$$= \log(p)\sum_t x^t + \log(1-p)\left(N - \sum_t x^t\right)$$
- By differentiating the above equation, the MLE of parameter p is given as
$$\hat{p} = \frac{\sum_t x^t}{N}$$

4 ML Estimation under Multinomial Distribution

- It is basically an extension to Binomial MLE.
- In this approach, we assume that K mutually exclusive states can be achieved with probability $p_i, i = 1, 2, \dots, K$ with $\sum_{i=1}^{K} p_i = 1$.
- However, ML estimate is same as in case of Binomial distribution.
$$\hat{p} = \frac{\sum_t x^t}{N}$$

5 ML Estimation under Normal Distribution

- If we assume the normal distribution, then the estimate of mean and variance is given as
$$\hat{\mu} = m = \frac{1}{N}\sum_t x^t$$
$$\widehat{\sigma^2} = s^2 = \frac{1}{N}\sum_t (x^t - m)^2$$

6 Bias, Variance and MSE of Normal Estimators

- Let $d = d(X)$ be the estimator of parameter θ.
- Then,
    - bias $b_\theta = E(d) - \theta$
    - MSE $e = E((d-\theta)^2)$
    - Let m represents sample mean, then,
        - $E(m) = \mu$
        - $Var(m) = {\sigma^2}/{N}$
    - Let s^2 represent estimate of sample variance
        - $E(s^2) = \frac{1}{N}\sum_t E((x^t)^2) - NE(m^2) = \frac{N-1}{N}\sigma^2$

7 Bayesian Estimator

- Estimation of parameter θ is done by posterior probability $P(\theta|X)$.
- MAP estimate of θ is $\underset{\theta}{\operatorname{argmax}}\, P(\theta|X)$.
- The optimal Bayesian is given as
$$\theta_B = \int \theta\, P(\theta|X) d\theta$$
- The mean of a normal distribution can be estimated as well. We assume $x^t \sim N(\mu, \sigma^2)$ and $\mu \sim N(\mu_0, \sigma_0^2)$. Then
$$E(\mu|X) = \frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\mu + \frac{\frac{1}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\mu_0$$

8 Parametric Classification

- Assuming we have K classes, the discriminant function is defined as $g_i(x) = P(x|C_i)P(C_i), i = 1, 2, \dots, K$. Instance x is classified as per the discriminant function which gives maximum value.

9 Multivariate Normal Distribution

- The probability density function is given as
$$P(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(D)}$$
Where D is Mahala Nabis distance $(x-\mu)^T\Sigma^{-1}(x-\mu)$
- Quadratic Discriminant function can be applied in this case. In this we assume that all features are independent and hence the covariance matrix is a diagonal matrix.
$$g_i(X) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - \mu_{ij}}{\sigma_{ij}}\right)^2 + \log(P(C_i))$$

10 Multivariate Discrete Features and Applications

- Attributes obey Bernoulli distribution.
- Discriminant is linear.
- Document characterization is one such application

11 Generalisation to Multinomial Cases

- Each x_j takes n_j discrete values.
- Dummy variables $z_{jk} = 1$ if $x_j = v_k$, else 0.
- Parameters are $p_{ijk} = P(z_{jk} = 1|C_i)$.
- Class likelihood = $\prod_{j=1}^{d}\prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$
- Discriminant: $\sum_j \sum_k z_{jk} \log(p_{ijk}) + \log(P(C_i))$

12 Non-parametric approaches

- Assumes similar inputs have similar outputs.
- Estimates probability density locally.
- Memory based learning.

13 Univariate Non-parametric density estimation

- Estimated probability density:
$$P(x) = [(\#(xt < x + h) - \#(xt < x))/N]/h$$
- Naive estimator:
$$P(x) = [(\#(xt < x + h/2) - \#(xt < x - h/2))/N]/h$$
- Can also use histogram of bin-width h

14 Kernel Estimator

- Kernel function: function of distance to determine weight of each sample
- Parzen Window:
$$P(x) = \frac{1}{Nh}\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)$$
- K(u)= 1 if |u|<½, else 0
- K(u) if normal then smooth estimator
- k-NN estimator:
- Adaptive kernel estimator
$$\frac{1}{N2d_k(x)}\sum_{t=1}^{N} K\left(\frac{x - x^t}{2d_k(x)}\right)$$

15 Multivariate Density Estimation

- Gaussian kernel is used in this case.
- For discrete: Hamming distance may be used.

16 Instance-based Learning

- Training: Store instances
- Testing: Retrieve and classify
- Compute locally, lazy learning

17 KNN Regression

- i^{th} neighbour of x: $\frac{\sum_{i=1}^{k} f(x_i)}{k}$
- Weight inversely proportional to square of distance, proportional to kernel function

18 Locally Weighted Regression

- $f(x) = w_0 + w_1x_1 + \dots + w_dx_d$
- MSE is computed in three scenarios.

Interesting Concepts

- KNN Regression

Novel Ideas out of lesson

- KNN Regression can be used in pattern recognition as well

Level of Preparation of 3rd Quiz

- Satisfactory