Sagnik Ghosh Homework I 3343-6044

1. Uniform distribution of a continuous variable x is

$$p(x;a,b) = \frac{1}{b-a} \qquad a \le x \le b$$

now $$\int_a^b \frac{1}{b-a}\,dx = \left|\frac{x}{b-a}\right|_a^b = \frac{b-a}{b-a} = 1$$

Therefore, the distribution is normalized.

& Mean = $$E(x) = \int_a^b \frac{x}{b-a}\,dx = \left|\frac{x^2}{2(b-a)}\right|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

$$Var(x) = E(x^2) - E^2(x)$$

Now, $$E(x^2) = \int_a^b \frac{x^2}{b-a}\,dx = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

So, $$Var(x) = \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4}$$

$$= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4}$$

$$= \frac{1}{12}(4b^2+4ab+4a^2-3a^2-6ab-3b^2)$$

$$= \frac{1}{12}(a^2-2ab+b^2)$$

$$= \frac{(a-b)^2}{12}$$

2. PMF of Poisson random variable,

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$ [likelihood of i.i.d samples]

taking log likelihood,

$$LL(\lambda) = \sum_{i=1}^{n} \left(\log \lambda^{x_i} + \log e^{-\lambda} - \log (x_i)!\right)$$

$$= \log\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \lambda - \sum_{i=1}^{n} \log (x_i)!$$

now for MLE,

$$\frac{\delta LL(\lambda)}{\delta(\lambda)} = \frac{1}{\lambda}(x_1+x_2+\ldots x_n) - n = 0$$

now $$\boxed{\lambda = \frac{x_1+x_2+\ldots+x_n}{n}}$$

3. random(d,1) generates multivariate normal $x \in R^d$ with 0 mean and I covariance.

therefore $x \sim \mathcal{N}(0, I)$

Applying affine transformation on x, ~~tis~~ in the form $Ax+b$.

From property of normal distribution,

$$Ax + b \sim \mathcal{N}(b, AA^T) \quad —(1)$$

Our goal is to generate random variable from $\mathcal{N}(\mu, \Sigma)$ —(2)

Comparing (1) and (2)

$$\mu = b.$$
$$\Sigma = AA^T.$$

Therefore to generate the required distribution, we need to perform Affine transformation on x, $(Ax + \mu)$.

$\mu \longrightarrow$ given as input.

$A \longrightarrow$ we can derive from Σ input.

Note, Σ is a covariance matrix, positive-definite matrix.

We can perform Cholesky Factorization to find matrix A.

4. $$\underset{\theta}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} r_i (y_i - \phi_i^T\theta)^2$$

This can be thought of as weighted least square.

The ordinary least square, $r_i \ \forall\, i \in \{1, n\} = 1$

Now, we write matrix R, with r_i on the diagonal. This could be rewritten,

$$WMSE = n^{-1} (Y - \phi^T\theta)^T R (Y - \phi^T\theta)$$

$$= \frac{1}{n}\left(Y^T R y - y^T R \phi^T\theta - \theta^T \phi R y + \theta^T \phi R \phi^T \theta\right)$$

Differentiating in terms of θ,

$$\nabla_y WMSE = \frac{2}{n} \left(-\phi R y + \phi R \phi^T \theta\right)$$

setting this to zero,

$$\hat{\theta} = (\phi R \phi^T)^{-1} \phi R y.$$

5.a) Naive Bayes,

a document D, whose class is given by C. So we choose the class C, which has the highest posterior Probability.

$$P(C|D) = \frac{P(D|C)\,P(C)}{P(D)} \propto P(D|C)\,P(C)$$

Now in bernoulli distribution, to calculate the likelihood of the data, we do not consider the frequency of the word, we only consider the availability of data.

$$P(D|C) = P(w_1|C) \ldots\ldots P(w_m|C)$$

① We would take the label and data for train to calculate the prior probability. P(C) gives us the number of documents percentage.

② For each class, we would calculate the occurence of a word across the documents. If any word (w_i) belongs to 100 documents of class (C_i) and if that class has total 200 documents, $P(w_i|C_i) = \frac{1}{2}$ then the likelihood $P(D|C) = \prod_{i=1}^{m} P(w_i|C)$

③ Now for ~~training~~ testing, we use the data generated from ① and ② to predict. If a document is present, we would use the probability of appearance and if it's absent we would consider the probability of not occuring.

$$f(D_j) = \underset{C}{\mathrm{argmax}} \prod_{i=1}^{m} \left[b_{ji}\, P(w_i|C) + (1-b_{ji})(1-P(w_i|C)\right] * P(C)$$

b_{ji} → whether ith word is present in document j

Accuracy = 62.33 %

b) Similar approach to bernoulli, however while calculating likelihood and prior, we are ~~not~~ going to consider the frequencies as well.

① Prior probability $P(C=K) = \frac{N_K}{N}$; if there are N documents in the training set and N_K belongs to class K

② Estimate the likelihoods $P(w_j|C=K) = \frac{\sum_{i=1}^{N} x_{ij} Z_{iK}}{\sum_{s=1}^{V} \sum_{i=1}^{N} x_{is} Z_{iK}}$.

$V \rightarrow$ n.o of words in histogram

$Z_{iK} = 1$, when D_i has class $C=K$.

③ Test data,

$$f(D_j) = \underset{c}{\operatorname{argmax}}\left(\sum_{i=1}^{m}(x_{ji} \log P(w_i|c)) + \log P(c)\right)$$

$m \longrightarrow$ n.o of words in histogram.

$x_{1j}, \ldots x_{1m} \longrightarrow$ count of occurrences of $x_1 \ldots x_m$ in j th document. of test data. $\rightarrow$ accuracy 77.76%

c) $P(C|D) \propto P(D|C).P(C)$ | $P(D|C)$ having normal distribution.

D is a document vector (x) | Assuming univariate normal distribution

$$p(D|C) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)\right)$$

$d \rightarrow$ length of histogram

so,

$$P(C|D) = \pi_c \exp(-\tfrac{1}{2}(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)) \cdot P(c)$$

now, $f(x) = \underset{c}{\operatorname{argmax}} \exp\left(-\tfrac{1}{2}(x-\mu_c)^T\Sigma^{-1}(x-\mu_c)\right) \cdot P(c)$

taking log,

$$= \underset{c}{\operatorname{arg\,min}}\ (x-\mu_c)^T\Sigma^{-1}(x-\mu_c) + \underset{c}{\operatorname{argmax}}\ P(c)$$

$$= \underset{c}{\operatorname{arg\,min}}\ x^T\Sigma^{-1}x + \mu_c^T\Sigma^{-1}\mu_c - 2\mu_c^T\Sigma^{-1}x + b_c$$

$$= \underset{c}{\operatorname{argmax}}\ (\mu_c^T\Sigma^{-1}x + b_c)$$

$\mu, \Sigma \longrightarrow$ We will get from TF-IDF, we will use the parameters to estimate $f(x)$ using