Mark Ghazvan
mg3627
ELEN 4903
Homework 1
Due Feb 5

Problem 1. Part 1

Imagine we have a sequence of N observations $(x_1, \ldots, x_N)$, where each $x_i \in \{0, 1\}$. We model this sequence as i.i.d Bernoulli random variables, where $p(x_i = 1 | \pi) = \pi$ and π is unknown

a) What is the joint likelihood of the data $(x_1, \ldots, x_N)$?

b) Derive the maximum likelihood estimate

$\hat{\pi}_{ML}$ for π.

c) Explain why this maximum likelihood estimate makes intuitive sense.

Solution

a) $p(x_1, \ldots, x_N | \pi) = \prod_{i=1}^{N} p(x_i | \pi)$

The probability ~~density~~ mass function can be written as $p(k | \pi) = \pi^k (1-\pi)^{1-k}$ $k \in \{0, 1\}$

Thus the joint likelihood function is

$$p(x_1, \ldots, x_N | \pi) = \prod_{i=1}^{N} \pi^{x_i} (1-\pi)^{1-x_i} =$$

$$= \pi^{\sum_{i=1}^{N} x_i} (1-\pi)^{N - \sum_{i=1}^{N} x_i}$$

Scanned by CamScanner

b) We have to solve the equation

$$\sum_{i=1}^{N} \frac{d}{d\pi} \ln\left(\pi^{\sum_{i=1}^{N} x_i} (1-\pi)^{N-\sum_{i=1}^{N} x_i}\right) =$$

$$= \sum_{i=1}^{N} \frac{d}{d\pi}\left(\sum_{i=1}^{N} x_i \ln\pi + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1-\pi)\right) =$$

$$= \sum_{i=1}^{N} \frac{1}{\pi}\sum_{i=1}^{N} x_i - \frac{1}{1-\pi}\left(N - \sum_{i=1}^{N} x_i\right)) =$$

$$= \sum_{i=1}^{N} x_i \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) - \frac{1}{1-\pi} N = 0$$

$$\left(\sum_{i=1}^{N} x_i\right)\left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) = \frac{1}{1-\pi} N$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{1-\pi}\frac{\pi(1-\pi)}{1} = \pi$$

So, $\hat{\pi}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$

c) Since $\sum_{i=1}^{N} x_i$ is the total number of number of successes observed in the ~~N~~ trials, $\hat{\pi}$ is the observed proportion of successes in the N trials or the average (mean) of successes observed.

Scanned by CamScanner

Part 2

Now imagine another sequence of N observations $(x_1, \ldots, x_N)$, where each $x_i \in \{0, 1, 2, \ldots\}$. We model this sequence as i.i.d. Poisson random variables with unknown parameter λ. The following questions follow exactly from Part I

a) What is the joint likelihood of the data $(x_1, \ldots, x_N)$?

b) Derive the max. likelihood estimate $\hat{\lambda}_{ML}$ for λ

c) Explain why this maximum likelihood estimate makes intuitive sense

Solution

a) The probability mass function of Poisson distribution $p(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

Thus $p(x_1, \ldots, x_N | \lambda) = \prod_{i=1}^{N} p(x_i | \lambda) =$

$$= \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{N} x_i} e^{-\lambda N}}{\prod_{i=1}^{N} x_i!}$$

b) $$\frac{d}{d\lambda} \ln\left(\frac{\lambda^{\sum_{i=1}^{N} x_i} e^{-\lambda N}}{\prod_{i=1}^{N} x_i!} \right) =$$

Scanned by CamScanner

$$= \frac{d}{d\lambda}\left(\ln\left(\lambda^{\sum_{i=1}^{N} x_i}\right) + \ln\left(e^{-\lambda N}\right) - \ln\left(\prod_{i=1}^{N} x_i!\right)\right) = \frac{d}{d\lambda}\left(\sum_{i=1}^{N} x_i \ln\lambda + -\lambda N - \sum_{i=1}^{N} \ln(x_i!)\right) =$$

$$= \frac{1}{\lambda}\sum_{i=1}^{N} x_i - N = 0$$

Hence $\hat{\lambda}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$

d) Since the Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of (from Wikipedia) and λ is the mean of expected value we can suppose that the average of observed values will be mean of the joint distribution.

Scanned by CamScanner

Problem 2

Again imagine we have a sequence of N non-negative integer-valued observations $(x_1, \ldots, x_n)$, which we model as i.i.d. Poisson random variables with unknown parameter λ. We place a gamma prior distribution on λ, written $\lambda \sim \text{Gam}(\lambda | a, b)$, where

$$\text{Gam}(\lambda | a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

a) Use Bayes rule to derive the posterior distribution of λ and identify the name of this distribution.

b) What is the mean and variance of λ under this posterior? Discuss how this relates to your solution to Part 2 of Problem 1.

Solution

a) $(x_1, \ldots, x_n)$ $x_i \sim \text{Poisson}(x_i | \lambda)$

$\lambda \sim \text{Gam}(\lambda | a, b)$

The posterior distribution of the parameter λ is $p(\lambda, x_1, \ldots, x_n)$

As we found at the previous problem

$$p(x_1, \ldots, x_n | \lambda) = \frac{\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!} e^{-\lambda n}$$

Scanned by CamScanner

The prior probability density function of λ is $\frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$.

The posterior distribution of λ, we can get from the Bayes rule:

$$p(\lambda / x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n / \lambda) p(\lambda)}{\int_0^\infty p(x_1, \dots, x_n / \lambda) p(\lambda) d\lambda} =$$

$$= \frac{\frac{b^a}{\Gamma(a)} \frac{1}{\prod_{i=1}^n x_i!} \lambda^{a-1+\sum x_i} e^{-\lambda(b+n)}}{\frac{b^a}{\Gamma(a)} \frac{1}{\prod x_i!} \int_0^\infty \lambda^{a-1+\sum x_i} e^{-\lambda(b+n)} d\lambda}$$

$$\int_0^\infty \lambda^{a-1+\sum x_i} e^{-\lambda(b+n)} d\lambda = \left\{ x = \lambda(b+n) \right\} =$$

$$= \int_0^\infty \left(\frac{x}{b+n} \right)^{a+\sum x_i - 1} e^{-x} \frac{1}{b+n} dx =$$

$$= \frac{1}{(b+n)^{a+\sum x_i}} \int_0^\infty x^{a+\sum x_i - 1} e^{-x} dx =$$

$$= \frac{1}{(b+n)^{a+\sum x_i}} \Gamma(a + \sum x_i)$$

As a result

$$p(\lambda / x_1, \dots, x_n) = \frac{(b+n)^{a+\sum x_i}}{\Gamma(a+\sum x_i)} \lambda^{a+\sum x_i - 1} e^{-\lambda(b+n)}$$

Scanned by CamScanner

This is a Gamma distribution with parameters $a + \sum_{i=1}^{n} x_i$, $b + n$

The posterior distribution of λ

$$p(\lambda | x_1, \ldots, x_n) = \frac{(b+n)^{a+\sum x_i}}{\Gamma(a+\sum x_i)} \lambda^{a+\sum x_i - 1} e^{-\lambda(b+n)}$$

b) The mean of a gamma distribution Gam (a, b) is equal to $\frac{a}{b}$, the variance $\frac{a}{b^2}$

Thus, $X \sim \theta$ Gam$(a + \sum_{i=1}^{n} x_i, b+n)$

has mean equal to $E(X) = \frac{a + \sum x_i}{b+n}$

and variance is equal to $\frac{a + \sum x_i}{(b+n)^2}$

Thus $\hat{\lambda} = \frac{a + \sum x_i}{b+n}$ is the Bayes estimates under MSE. From part 2 of Problem 1 we see that the a mean is the best estimation of a parameter of the Poisson distribution and the estination $\hat{\lambda}$ is simply the mean of the posterior distribution.

Scanned by CamScanner

Part 1. First, randomly split the data set into 20 testing examples and 372 training examples. Using the training data only, solve a linear regression model of the form $y \approx w_0 + \sum_{j=1}^{6} x_j w_j$ using least squares.

(a) Print the numbers you obtain for the vector $\hat{w}_{\text{ML}}$. Using the labels of each dimension contained in the readme file, explain what the sign of each value in $\hat{w}_{\text{ML}}$ says about the relationship of the inputs to the output.

We get the next w:
w0=23.4125374120238
w1=-0.351950426082283
w2=0.219371403547162
w3=0.0290871251139913
w4=-5.52965644605478
w5=0.153587095845352
w6=2.70194113867137

y: miles per gallon
- x1: intercept term
- x2: number of cylinders
- x3: displacement
- x4: horsepower
- x5: weight
- x6: acceleration
- x7: model year

Thus we see that number of cylinder and weight reduce consumption of patrol (more cylinders or weight – less consumption, but it shows only correlation). On the other hand when displacement, horsepower, acceleration and model year go up consumption goes up too.

Intercept term is a basic level of consumption.

(b) Use the least squares solution to predict the outputs for each of the 20 testing examples. Repeat this process of randomly splitting into training and testing sets 1000 times. Each time, calculate the mean absolute error of the resulting predictions, $\text{MAE} = \frac{1}{20} \sum_{i=1}^{20} |y_i^{\text{test}} - y_i^{\text{pred}}|$. What is the mean and standard deviation of the MAE for these 1000 tests?

We get the mean

MAE_mean = 2.4979

We get the standard derivation

MAE_std = 0.3797

Part 2. Using exactly the same training/testing setup as in Part 1, fit a pth order polynomial regression model using least squares for $p = 1, 2, 3, 4$. (Note that $p = 1$ is equivalent to Part 1.) For each value of p run 1000 experiments on randomly partitioned training/testing sets using 20 testing and 372 training examples. For each experiment calculate the root mean squared error,

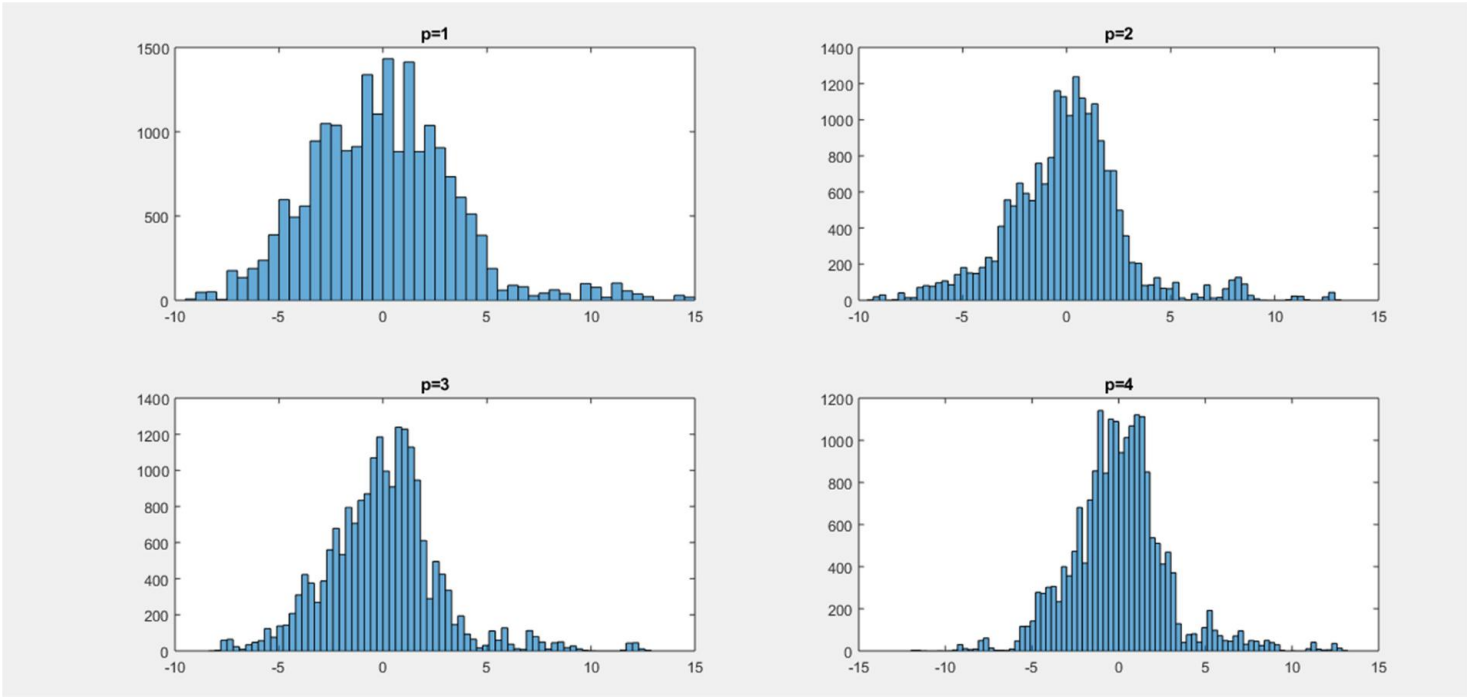
$$\text{RMSE} = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(y_i^{\text{test}} - y_i^{\text{pred}})^2}.$$

(a) In a table, print the mean and standard deviation of the RMSE as a function of p. Using these numbers argue for which value of p is the best.

p	1	2	3	4
RMSE_mean	3.4219	2.7628	2.6431	2.7303
RMSE_std	0.6738	0.6428	0.6027	0.6372

As we can see p=3 is the best one, since it has the least mean and the least variation.

(b) For each value of p, collect $y^{\text{test}} - y^{\text{pred}}$ for each test example. (Observe that this number can be negative, and there are 20×1000 in total.) Plot a histogram of these errors for each p.



(c) For each p, use maximum likelihood to fit a univariate Gaussian to the 20,000 errors from Part 2(b). Describe how you calculated the maximum likelihood values for the mean and variance (this is a univariate case of what we did in class, so no need to re-derive it). What is the log likelihood of these empirical errors using the maximum likelihood values for the mean and variance? Show this as a function of p and discuss how this agrees/disagrees with your conclusion in Part 2(a). What assumptions are best satisfied by the optimal value of p using this approach?

We calculate the mean and variance in the following way:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

And we calculate the log likelihood as

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}$$

.

And we get

p	1	2	3	4
Error _mean	-0.0063	0.0133	-0.0036	0.0203
Error_std	3.4876	2.8366	2.7109	2.8036
Log likelihood	-2.0885e+04	-1.9362e+04	-1.8892e+04	-1.9211e+04

As we can see the largest log likelihood occurs when p=3 that agrees with our conclusion in Part 2(a). We are making an independent Gaussian noise assumption about the error with parameters mean = -0.0036 (but we had to get 0) and standard deviation =2.7109.