SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Answer #1

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *September 25th, 2019*

Question 1

Consider the polynomial function

$$y(x, \mathrm{w}) = w_0 + w_1 x + w_2 x + ... + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Calculate the coefficients $\mathrm{w} = \{w_i\}$ that minimize its sum-of-squares error function. Here a suffix i or j denotes the index of a component, whereas $(x)^i$ denotes x raised to the power of i.

Answer. Submit $y(x, \mathrm{w})$ into

$$E(\mathrm{w}) = \frac{1}{2} \sum_{n=1}^{N} y(x_n, \mathrm{w} - t_n)^2$$

and then differentiating with w_i we obtain

$$\sum_{n=1}^{N} (\sum_{j=0}^{M} + \mathrm{w}_j x_n^j - t_n) x_n^i = 0.$$

Re-arranging terms and then we can obtain

$$\sum_{j=0}^{M} A_{ij} \mathrm{w}_j = T_i,$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n.$$

Question 2

Suppose that we have three colored boxes r(red), b(blue), and g(green).Box r contains 3 apples, 4 oranges, and 3 limes, box $b$contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2, p(b) = 0.2, p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Answer. Let us denote appales, oranges and limes by a, o and l respectively. The marginal probability of selecting an apple is given by

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g)$$

$$= \frac{3}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.34$$

where the conditional probabilities are obtained from the proportions of apples in each box.

To find the probability that the bos was green, given that the fruit we selected was orange, we can use Bayes' theorem

$$p(g|o) = \frac{p(o|g)p(g)}{p(o)}.$$

The denominator is given by

$$p(o) = p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g)$$

$$= \frac{4}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.36$$

from which we obtain

$$p(g|o) = \frac{3}{10} \times \frac{0.6}{0.36} = \frac{1}{2}.$$

Question 3

Given two statistically independent variables x and z, show that the mean and variance of their sum satisfies

$$\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z]$$

$$\mathrm{var}[x+z] = \mathrm{var}[x] + \mathrm{var}[z]$$

Answer. Since x and z are independent, their joint distribution factories $p(x,z) = p(x)p(z)$, and so

$$\begin{aligned}\mathbb{E}[x+z] &= \int\int (x+z)p(x)p(z)\mathrm{d}x\mathrm{z} \\ &= \int xp(x)\mathrm{d}x + \int zp(z)\mathrm{d}z \\ &= \mathbb{E}[x] + \mathbb{E}[z].\end{aligned}$$

Similarly for variances, we first note that

$$(x+z-\mathbb{E}[x+z])^2 = (x-\mathbb{E}[x])^2 + (z-\mathbb{E}[z])^2 + 2(x-\mathbb{E}[x])(z-\mathbb{E}[z])$$

where the final term will integrate to zero with respect to the fractorized distribution $p(x)p(z)$.
Hence

$$\begin{aligned}\mathrm{var}[x+z] &= \int\int (x+z-\mathbb{E}[x+z])^2 p(x)p(z)\mathrm{d}x\mathrm{d}z \\ &= \int (x-\mathbb{E}[x])^2 p(x)\mathrm{d}x + \int (z-\mathbb{E}[z])^2 p(z)\mathrm{d}z \\ &= \mathrm{var}(x) + \mathrm{var}(z).\end{aligned}$$

For discrete variables the integrals are replaced by summations, and the same results are again obtained.

Question 4

In probability theory and statistics, the Poisson distribution, is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. If X is Poisson distributed, i.e. $X \sim Possion(\lambda)$, its probability mass function takes the following form:

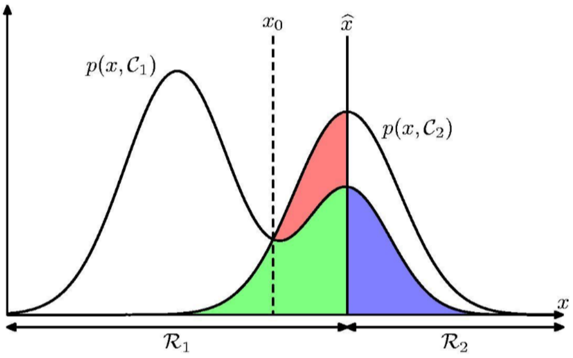
$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

It can be shown that if $\mathbb{E}(X) = \lambda$. Assume now we have n data points from $Possion(\lambda) : \mathcal{D} = \{X_1, X_2, ..., X_n\}$. Show that the sample mean $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the maximum likelihood estimate(MLE) of λ. If X is exponential distribution and its distribution density function is $f(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. Show that the sample mean $\widehat{\lambda}\frac{1}{n}\sum_{i=1}^{n} X_i$ is the maximum likelihood estimate(MLE) of λ.

Answer. Waiting...

Question 5

(a) Write down the probability of classifying correctly $p(correct)$ and the probability of misclassification $p(mistake)$ according to the following chart.



(b) For multiple target variables described by vector $\mathbf{t}$, the expected squared loss function is given by

$$\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))] = \int\int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}$$

Show that the function $\mathbf{y}(\mathbf{x})$ for which this expected loss is minimized given by $\mathbf{y}(\mathbf{x}) = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$.

Hints. For a single target variable t, the loss is given by

$$\mathbb{E}[L] = \int\int y(\mathbf{x}) - t^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}t$$

The result is as follows

$$y(\mathbf{x}) = \frac{\int tp(\mathbf{x}, t)\mathrm{d}t}{p(\mathbf{x})} = \int tp(t|\mathbf{x}\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}])$$

Answer. (a)

$$p(correct) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_{\mathbf{k}}, \mathcal{C}_{\mathbf{k}}) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_{\mathbf{k}})\mathrm{d}\mathbf{x}.$$

$$\begin{aligned} p(mistake) &= p(\mathbf{x} \in \mathcal{R}_{\mathbf{1}}, \mathcal{C}_{\mathbf{2}}) + p(\mathbf{x} \in \mathcal{R}_{\mathbf{2}}, \mathcal{C}_{\mathbf{1}}) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_{\mathbf{2}})\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_{\mathbf{1}})\mathrm{d}\mathbf{x} \end{aligned}$$

(b) Our goal is to choose $y(x)$ so as to minimize $\mathbb{E}[L]$. We can do this formally using the calculus of variations to give

$$\frac{\delta \mathbb{E}[L]}{\delta \mathbf{y}(\mathbf{x})} = \int 2(\mathbf{y}(\mathbf{x}) - \mathbf{t})p(\mathbf{t}, \mathbf{x})\mathrm{d}\mathbf{t} = 0.$$

Solving for $\mathbf{y}(\mathbf{x})$, and using the sum and product rules of probability, we can obtain

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t}p(\mathbf{t},\mathbf{x})\mathrm{d}\mathbf{t}}{\int p(\mathbf{t},\mathbf{x})\mathrm{d}\mathbf{t}} = \int \mathbf{t}p(\mathbf{t}|\mathbf{x})\mathrm{d}\mathbf{t}$$

which is the conditional average of **t** conditioned on **x**.

Question 6

(*a*) We defined the entropy based on a discrete random variable **X** as

$$\mathbf{H}[\mathbf{X}] = -\sum_i p(x_i)\ln p(x_i)$$

Now consider the case that **X** is a continuous random variable with the probability density function $p(x)$. The entropy is defined as

$$\mathbf{H}[\mathbf{X}] = -\int p(x)\ln p(x)dx$$

Assume that **X** follows Gaussian distribution with the mean μ and variance σ, i.e.

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

Please derive its entropy $\mathbf{H}[\mathbf{X}]$.

(*b*) For a density defined over multiple continuous variables, denoted collectively by the vector **x**, write down the differential entropy $\mathbf{H}[\mathbf{x}]$, the conditional entropy $\mathbf{H}(\mathbf{y}|\mathbf{x})$ and the mutual information $\mathbf{I}(\mathbf{y}|\mathbf{x})$. Then show the following equation

$$\mathbf{I}[\mathbf{x},\mathbf{y}] = \mathbf{H}[\mathbf{x}] - \mathbf{H}[\mathbf{x}|\mathbf{y}] = \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}]$$

Answer. (a) Waiting...

(b)

$$\begin{aligned}\mathbf{I}[\mathbf{x},\mathbf{y}] &\equiv KL(p(\mathbf{x},\mathbf{y}) \parallel p(\mathbf{x})p(\mathbf{y})) \\ &= -\int\int p(\mathbf{x},\mathbf{y})\ln(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\end{aligned}$$

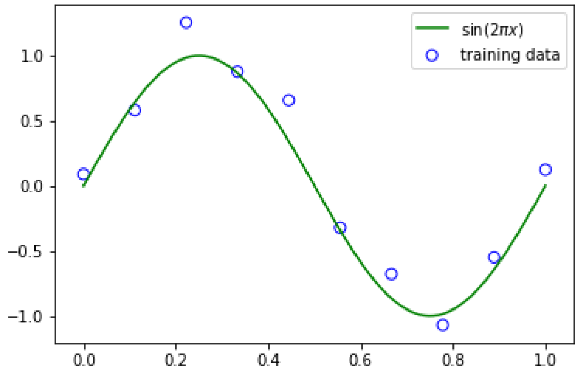
From the product rule we have $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$, so

$$\begin{aligned}\mathbf{I}[\mathbf{x},\mathbf{y}] &= -\int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}))\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x}))\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y})\ln p(\mathbf{y}))\mathrm{d}\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x}))\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &= \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}]\end{aligned}$$

Program Question

You should download the `HW1_programQuestion.ipynb` file first.

(*a*) Plot the graph with given code, the result should be same as this.



(*b*) On the basis of the results, you should try 0^{th} order polynomial, 1^{st} order polynomial, 3^{rd} order polynomial and some other order polynomial, show the results include fitting and over-fitting.

(*c*) Plot the graph of the root-mean-square error.

(*d*) Plot the graph of the predictive distribution resulting from a Bayesian treatment of polynomial curve fitting using an M=9 polynomial, with the fixed parameters $\alpha = 5 \times 10^{-3}$ and $\beta = 11.1$(corresponding to the known noise variance).

(*e*) Change the *sample_size* to 2, 3 or 10 times than before, explain the change of M.

Hints. You should install `matplotlib.pyplot`, and read classes `PolynomialFeature`, `LinearRegression`, and `BayesianRegression` in the file.