Intruduction to Machine Learning

Dr S.Amini

دانشگاه صنعتی شریف

department: Electrical Engineering

Amirreza Velae 400102222
github repository

Homework 1

April 17, 2023

Intruduction to Machine Learning

Homework 1

Amirreza Velae 400102222
github repository

Correlation, Causality, and Independence

Let X ∼Uniform$(-1, 1)$, and $Y = X^2$. Clearly, X and Y aren't independent. (Actually , they have a causation property!). Show that even though they are dependant, they are uncorrelated, which means $X, Y = 0$.

solution

to show that two random variables are dependant, we need to show that they are not independent.First we observe Y's distribution.

$$F_x(x) = \frac{x+1}{2} \;\;\&\;\; f_X(x) = \frac{1}{2} \;\; \forall x \in [-1, 1] \quad \& \quad f_Y(y) = \sum_{x:h(x)=y} \frac{f_X(x)}{|h'(x)|}$$

$$\implies f_Y(y) = \frac{f_{X_1}(x_1)}{|\frac{\mathrm{d}\sqrt{x_1}}{\mathrm{d}x_1}|} + f_Y(y) = \frac{f_{X_2}(x_2)}{|\frac{\mathrm{d}\sqrt{x_2}}{\mathrm{d}x_2}|}, \qquad x_1 = \sqrt{y}, \;\; x_2 = -\sqrt{y}$$

$$\implies f_Y(y) = \frac{1}{2|2\sqrt{y}|} + \frac{1}{2|-2\sqrt{y}|} = \frac{1}{2\sqrt{y}} \;\; \forall y \in [0, 1]$$

$$P(Y = y, X = x) = P(Y = x^2, X = x)$$

$$\neq P(Y = y)P(X = x) = \frac{1}{4\sqrt{y}} \;\; \forall y \in [0, 1] \;\; \& \;\; \forall x \in [-1, 1]$$

Now we show that X and Y are not correlated; that is, the correlation coefficient is zero (i.e., $\rho_{X,Y} = 0$).To show that X and Y are not correlated, we need to show that covariance is zero(i.e.,$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = 0$).

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$$

$$E[XY] = E[E[XY|X]] = E[XE[Y|X]] = E[X^3] = \int_{-1}^{1} x^3 \frac{1}{2} dx = 0$$

$$E[X] = \int_{-1}^{1} x\frac{1}{2}dx = 0 \;\;,\;\; E[Y] = \int_0^1 y\frac{1}{2\sqrt{y}}dy = \int_0^1 \frac{y}{\sqrt{y}}dy = \frac{2}{3}$$

$$\implies Cov(X, Y) = 0 \implies \rho_{X,Y} = 0$$

Markov-Chain Gaussians

We write $X \to Y \to Z$ and say that X, Y, and Z form a Markov chain when we have: $X|Y \perp Z|Y$ which also means $P_{X,Z|Y}(z,x|y) = P_{X|Y}(x|y)P_{Z|Y}(z|y)$. For three Gaussians variables with the preceding property, compute $\rho_{X,Z}$ in terms of $\rho_{X,Y}$ and $\rho_{Y,Z}$.

solution

Correlation covariance is defined as: $\rho_{X,Y} = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}}$ .So:

$$\rho_{X,Z} = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}} = \frac{E[XY] - E[X]E[Y]}{\sqrt{Var(X)Var(Z)}} = \frac{E[E[XZ|Y]] - E[X]E[Z]}{Var(X)Var(Z)}$$
$$\frac{E[E[X|Y]E[Z|Y]] - E[X]E[Z]}{\sigma_X \sigma_Y}$$

From the formula that is given in the HW:

$$E[X|Y] = \mu_X + \frac{Cov(X,Y)}{Var(Y)}(Y - \mu_Y) = \mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y)$$
$$E[Z|Y] = \mu_Z + \frac{Cov(Z,Y)}{Var(Y)}(Y - \mu_Y) = \mu_Z + \rho_{Z,Y}\frac{\sigma_Z}{\sigma_Y}(Y - \mu_Y)$$
$$\implies E[E[X|Y]E[Z|Y]] = E[(\mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y))(\mu_Z + \rho_{Z,Y}\frac{\sigma_Z}{\sigma_Y}(Y - \mu_Y))]$$
$$= \mu_X\mu_Z + E[\mu_X\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y) + \mu_Z\rho_{Z,Y}\frac{\sigma_Z}{\sigma_Y}(Y - \mu_Y) + \rho_{X,Y}\rho_{Z,Y}\frac{\sigma_X}{\sigma_Y}\frac{\sigma_Z}{\sigma_Y}(Y - \mu_Y)^2]$$
$$= \mu_X\mu_Z + (\mu_X\rho_{X,Y}\frac{\sigma_X}{\sigma_Y} + \mu_Z\rho_{Z,Y}\frac{\sigma_Z}{\sigma_Y})(E[Y] - \mu_Y) + \rho_{X,Y}\rho_{Z,Y}\frac{\sigma_X}{\sigma_Y}\frac{\sigma_Z}{\sigma_Y}Vay(Y)$$
$$= \mu_X\mu_Z + \rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z$$

So:

$$\rho_{X,Z} = \frac{E[E[X|Y]E[Z|Y]] - E[X]E[Z]}{\sigma_X\sigma_Y} = \frac{\mu_X\mu_Z + \rho_{X,Y}\rho_{Z,Y}\sigma_X\sigma_Z - \mu_X\mu_Z}{\sigma_X\sigma_Y} = \rho_{X,Y}\rho_{Z,Y}$$

Sensor Fusion

Imagine the temperature is a fixed number z (which we know nothing about. You can model it with $Z \sim \mathcal{N}(0, +\infty)$). We have two sensors, in which the temperature is measured with noise. The variance of noise for each of them is known and it's v_1 and v_2 respectively. Suppose we make n_1 observation from the first sensor, each given by $\{Y_1^{(i)}\}_{i=0}^{n_1}$ and n_2 observation of the second sensor given by $\{Y_2^{(i)}\}_{i=0}^{n_2}$. Consider all of these observations to be shown as a set called $\mathcal{D}$, Using the given variances, find$p_{Z|\mathcal{D}}(z|\mathcal{D})$ and estimate Z using its mean.

solution

Assume the following items:

- $\mathbb{Z} \in \mathbb{R}^L$:Unknown vector
- $\mathbb{Y} \in \mathbb{R}^D$:Noisy measurements
- The following distributions hold:
  $p(z) = \mathcal{N}(z|\mu_z, \Sigma_z)$
  $p(y|z) = \mathcal{N}(y|Wz + b, \Sigma_y), W \in \mathbb{R}^{D\times L}, b \in \mathbb{R}^D$

Then:

- Joint distribution $p(z, y) = p(z)p(y|z)$ is a $L + D$ dimensional Gaussian with the following parameters:

$$\mu_= \begin{bmatrix} \mu_z \\ W\mu_z + b \end{bmatrix},\ \Sigma_= \begin{bmatrix} \Sigma_z & \Sigma_z W^T \\ W\Sigma_z & \Sigma_y + W\Sigma_z W^T \end{bmatrix}$$

- Using Bayes rule, the posterior $p(z|y)$ is also L is also L dimensional Gaussian with the following parameters:

$$\Sigma_{z|y}^{-1} = \Sigma_z^{-1} + W^T\Sigma_y^{-1}W$$
$$\mu_{z|y} = \Sigma_{z|y}[W^T\Sigma_y^{-1}(y - b) + \Sigma_z^{-1}\mu_z]$$

$$\mathcal{D} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \ \&\ y_1 = \begin{bmatrix} Y_1^{(0)} \\ \vdots \\ Y_1^{(n-1)} \end{bmatrix} \ \&\ y_2 = \begin{bmatrix} Y_2^{(0)} \\ \vdots \\ Y_2^{(n-1)} \end{bmatrix} \ \&\ W = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

The $\{Y_1^{(i)}\}_{i=0}^{n_1}$ and $\{Y_2^{(i)}\}_{i=0}^{n_2}$ are independent, therfore they are uncorrelated. So:

$$\Sigma_{\mathcal{D}} = \begin{bmatrix} \sigma_{Y_1}^2 & 0 \\ 0 & \sigma_{Y_2}^2 \end{bmatrix} = \begin{bmatrix} v_1 * I_{n_1\times n_1} & 0 \\ 0 & v_2 * I_{n_2\times n_2} \end{bmatrix} \implies \Sigma_{\mathcal{D}}^{-1} = \begin{bmatrix} \frac{1}{v_1} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \frac{1}{v_1} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{v_2} & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & \frac{1}{v_2} \end{bmatrix}$$

$$\implies \Sigma_{z|\mathcal{D}}^{-1} = \Sigma_z^{-1} + W^T\Sigma_{\mathcal{D}}^{-1}W = n_1 * \frac{1}{v_1} + n_2 * \frac{1}{v_2} = \frac{n_1}{v_1} + \frac{n_2}{v_2}$$

$$\mu_{z|\mathcal{D}} = \Sigma_{z|\mathcal{D}}[W^T\Sigma_{\mathcal{D}}^{-1}(\mathcal{D}-b)+\Sigma_z^{-1}\mu_z] = \frac{1}{\frac{n_1}{v_1}+\frac{n_2}{v_2}}\begin{bmatrix}1 & \dots & 1\end{bmatrix}\begin{bmatrix}\frac{1}{v_1} & 0 & \dots & 0 & 0\\ 0 & \frac{1}{v_1} & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & \frac{1}{v_2} & 0\\ 0 & 0 & \dots & 0 & \frac{1}{v_2}\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}$$

$$= \frac{v_1v_2}{n_1v_2+n_2v_1}\times\begin{bmatrix}\frac{1}{v_1} & \frac{1}{v_1} & \dots & \frac{1}{v_2} & \frac{1}{v_2}\end{bmatrix}\cdot\begin{bmatrix}Y_1^{(0)}\\ Y_1^{(1)}\\ \dots\\ Y_1^{(n_1-1)}\\ Y_2^{(0)}\\ Y_2^{(1)}\\ \dots\\ Y_2^{(n_2-1)}\end{bmatrix} = \frac{v_1v_2}{n_1v_2+n_2v_1}\times(\frac{Y_1^{(0)}}{v_1}+\dots+\frac{Y_2^{(1)}}{v_2})$$

$$= \frac{1}{n_1v_2+n_2v_1}[(n_1v_2)E[Y_1]+(n_2v_1)E[Y_2]] = \frac{1}{n_1v_2+n_2v_1}[v_2n_1\mu_{Y_1}+v_1n_2\mu_{Y_2}]$$

Pseudo Inverse

Assume that matrix A has an SVD decomposition $A = U\Sigma V^\top$. We define the pseudo inverse of A as $A^\dagger = V\Sigma^{-1}U^\top$. Prove the followings:

- if A has a full row rank, then $A^\dagger = A^\top(AA^\top)^{-1}$.

Soloution

$$A^\top = V\Sigma^\top U^\top \implies AA^\top = U\Sigma V^\top V\Sigma^\top U^\top = U\Sigma^2U^\top$$
$$A^\top(AA^\top)^{-1} = A^\top\ U\Sigma^{-2}U^\top = V\Sigma^\top U^\top\ U\Sigma^{-2}U^\top = V\Sigma^{-1}U^\top$$
$$\implies A^\dagger = V\Sigma^{-1}U^\top = A^\top(AA^\top)^{-1}$$

- if A has a full column rank, then $A^\dagger = (A^\top A)^{-1}A^\top$.

Soloution

$$A^\top = V\Sigma^\top U^\top \implies A^\top A = V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma^2V^\top$$
$$(A^\top A)^{-1}A^\top = V\Sigma^{-2}V^\top\ V\Sigma^\top U^\top = V\Sigma^{-2}V^\top U\Sigma U^\top = V\Sigma^{-1}U^\top$$
$$\implies A^\dagger = V\Sigma^{-1}U^\top = (A^\top A)^{-1}A^\top$$

Eigenvalues

We show the eigenvalues of the square matrix A by $\lambda_1, \lambda_2, \dots, \lambda_n$. Prove the following:

- $$Tr\{A\} = \sum_{i=1}^{n} \lambda_i$$

Soloution

By spectral decomposition, we can write $\mathrm{Tr}(A = P'\Lambda P) = \mathrm{Tr}(\Lambda P P')$, where P is an orthogonal matrix and Λ is a diagonal matrix with the eigenvalues of A on the diagonal. Therefore, we have:

$$Tr\{A\} = Tr\{P'\Lambda P\} = Tr\{\Lambda P P'\} = Tr\{\Lambda\} = \sum_{i=1}^{n} \lambda_i$$

- $$det\{A\} = \prod_{i=1}^{n} \lambda_i$$

Soloution

$$\text{Write} A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

Also let the n eigenvalues of A be $\lambda_1, \lambda_2, \dots, \lambda_n$. Finally let the characteristic polynomial of A be $p_A(x) = \det(A - xI) = \lambda^n + c_{n-1}\lambda^{n-1} + \ldots + c_1\lambda + c_0$.
Note that since the eigenvalues of A are the zeros of $P(\lambda)$, this implies that $p_A(\lambda)$ can be factorized as $p_A(\lambda) = (\lambda - \lambda_1) \ldots (\lambda - \lambda_n)$.
Consider the constant term of $p(\lambda)$, c_0. The constant term of $p(\lambda)$ is given by $p(0)$, which can be calculated in the following two ways:

- $P(0) = (0 - \lambda_1) \ldots (0 - \lambda_n) = (-1)^n \lambda_1 \ldots \lambda_n$.
- $p(0) = \det(A - 0I) = \det(A) = \lambda_1 \ldots \lambda_n$.

Therefore, we have $Tr\{A\} = \sum_{i=1}^{n} \lambda_i$.

Maximum Likelihood Estimation

Suppose we have a random vector $X \in \mathbb{R}^d$. All elements are assumed to be iid random variables. Assume that we have an observation x. We want to fit a probability distribution to this data and we are going to use the maximum likelihood for that.

Bernoulli random variable

Assume that each X_i is a Bernoulli random variable, i.e., $p_{X_i}(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$. Also assume that we have observed m ones and k zeros. Find the distribution parameter θ.

Soloution

$$\hat{\theta}_{mle} = \underset{\theta}{argmax}\, p(x|\theta)$$

$$\log p(x|\theta) = \log \prod_{i=1}^{m} \theta + \log \prod_{i=1}^{k} (1-\theta)$$

$$= m \log \theta + k \log(1-\theta)$$

$$\implies \frac{\partial}{\partial \theta} \log p(x|\theta) = \frac{m}{\theta} - \frac{k}{1-\theta} = 0$$

$$\implies \theta = \frac{m}{m+k}$$

Exponential random variable

Assume that each X_i is a Exponential random variable, i.e., $p_{X_i}(x_i) = \lambda e^{-\lambda x_i} \mathbf{1}\{x_i \geq 0\}$. Also assume that all x_i values are positive. Find the exponential parameter λ.

Soloution

$$\hat{\lambda}_{mle} = \underset{\lambda}{argmax}\, p(x|\lambda)$$

$$\log p(x|\lambda) = \log \prod_{i=1}^{m} \lambda e^{-\lambda x_i}$$

$$= m \log \lambda - \lambda \sum_{i=1}^{m} x_i$$

$$\implies \frac{\partial}{\partial \lambda} \log p(x|\lambda) = \frac{m}{\lambda} - \sum_{i=1}^{m} x_i = 0$$

$$\implies \lambda = \frac{m}{\sum_{i=1}^{m} x_i}$$

Normal random variable

Assume that each X_i is a Normal random variable, i.e., $p_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$. Find the mean and variance of the distribution.

Soloution

$$\hat{\mu}_{mle} = \underset{\mu}{argmax}\, p(x|\mu)$$

$$\log p(x|\mu) = \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = -\frac{m}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2$$

$$\implies \frac{\partial}{\partial\mu}\log p(x|\mu) = -\frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu) = 0$$

$$\implies \mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\hat{\sigma}_{mle} = \underset{\sigma}{argmax}\, p(x|\sigma)$$

$$\log p(x|\sigma) = \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = -\frac{m}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2$$

$$\implies \frac{\partial}{\partial\sigma}\log p(x|\sigma) = \frac{m}{2\sigma^3} - \frac{1}{2\sigma^3}\sum_{i=1}^{m}(x_i-\mu)^2 = 0$$

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

A Tiny Bit of Vector Differentiation

Prove the following differentiation formulas. These formulas will be useful throughout the course.

- $$\boldsymbol{\nabla}_X(a^\top X) = [\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_d}]^\top (a^\top X) = a^\top$$

Soloution

$$\frac{\partial}{\partial x_i}(a^\top X) = \frac{\partial}{\partial x_i}\sum_{j=1}^{d} a_j x_j = a_i$$

$$\implies \boldsymbol{\nabla}_X(a^\top X) = [\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_d}]^\top (a^\top X) = a^\top$$

- $$\boldsymbol{\nabla}_X(X^\top AX) = [\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_d}]^\top (X^\top AX) = x^\top(A+A^\top) = (A+A^\top)X$$

Soloution

$$\frac{\partial}{\partial x_i}(X^\top AX) = \frac{\partial}{\partial x_i}\sum_{j=1}^{d}\sum_{k=1}^{d} x_j a_{jk} x_k = \sum_{k=1}^{d} a_{ik}x_k + \sum_{j=1}^{d} x_j a_{ji}$$

$$\text{Also:} \sum_{k=1}^{d} a_{ik}x_k + \sum_{j=1}^{d} x_j a_{ji} = \sum_{k=1}^{d} a_{ik}x_k + \sum_{j=1}^{d} a_{ij}x_j = \sum_{k=1}^{d} x_k a_{ik} + \sum_{j=1}^{d} x_j a_{ji}$$

$$\implies \boldsymbol{\nabla}_X(X^\top AX) = [\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_d}]^\top (X^\top AX) = x^\top(A + A^\top) = (A + A^\top)X$$

End of Homework 1