Mock Multiple Choice Questions — com4509/com6509

1. [probability] A random variable X takes a value x, drawn from a set $\mathcal{X}$ containing possible outcomes (or events). Let $p(x)$ denote the probability density function of X. The discrete type expected value of a random variable X is defined as

Slide 1

a) $E[X] = \sum_{x \in \mathcal{X}} \frac{d}{dx} p(x)$

b) $E[X] = \sum_{x \in \mathcal{X}} p(x)$

c) $E[X] = \sum_{x \in \mathcal{X}} x p(x)$

d) $E[X] = \sum_{x \in \mathcal{X}} x^2 p(x)$

e) non of the above

Expected Value

The **expected value** (or mean, average) of a random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x) dx$$

- discrete type is $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p(x)$ for all possible events $\mathcal{X}$

The expected value of a function $f(x)$ is

$$\mathbb{E}[f(x)] = \int_{-\infty}^{\infty} f(x) p(x) dx$$

2. [probability] A random variable X (or Y) takes a value x (or y), drawn from a set $\mathcal{X}$ (or $\mathcal{Y}$) containing possible outcomes. Which one of the following is the correctly formulated?

Slide 1

a) $p(x|y) = p(y|x)p(x)$

b) $p(x|y) = \frac{p(y|x)p(x)}{\sum_{x \in \mathcal{X}} p(y|x)p(x)}$

c) $p(x|y) = p(y|x)p(y)$

d) $p(x|y) = \frac{p(y|x)p(y)}{\sum_{x \in \mathcal{X}} p(y|x)p(y)}$

e) $p(x|y) = \frac{p(y|x)}{\sum_{x \in \mathcal{X}} p(y|x)}$

Product Rule and Sum Rule

The product rule of probability:

$$\underbrace{p(x, y)}_{\text{joint probability}} = \underbrace{p(y|x)}_{\text{conditional probability}} \cdot p(x)$$

The sum rule of probability:

$$\underbrace{p(y)}_{\text{marginal probability}} = \sum_{x \in \mathcal{X}} p(x, y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$$

Bayes' Theorem

Bayes' theorem immediately follows the product rule:

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(y|x)p(x)}{\sum_{x \in \mathcal{X}} p(y|x)p(x)}$$

3. [vector calculation] There are three vectors, $\mathbf{x} = (1, 3, -5)$, $\mathbf{y} = (4, -2, -1)$ and $\mathbf{z} = (-2, -5, 2)$. Which of the following is correct?

a) $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

b) $\mathbf{y}$ and $\mathbf{z}$ are orthogonal.

$\mathrm{y} \cdot \mathrm{z} = 0$

c) $\mathbf{x}$ and $\mathbf{z}$ are orthogonal.

d) $\mathbf{x}$ and $\mathbf{y}$ and $\mathbf{z}$ are orthogonal.

e) non of the above.

4. [multivariate normal distribution] Using an $n \times 1$ column vector μ and an $n \times n$ square matrix Σ, the multivariate normal distribution of an n-dimensional random vector $\mathbf{x}$ is

Slide 1

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \Sigma}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)\right\}$$

where $(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)$ is ← $1 \times n$, $n \times n$, $n \times 1 = 1 \times 1$ (scalar value)

a) $n \times n$ square matrix

b) $n \times 1$ column vector

c) $1 \times n$ row vector

d) scalar value

e) non of the above

Multivariate Gaussian (Normal) Distribution

The probability density of the k-dimensional **Gaussian distribution** is

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu}$ is the $k \times 1$ **mean vector** and $\boldsymbol{\Sigma}$ is the $k \times k$ **covariance matrix**.

- $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ are the **determinant** and the **inverse** of the covariance
- a symbol $^\top$ indicates the **transpose**

5. [overdetermined systems] Consider the following four overdetermined systems, each consisting of three equations and two unknowns x and y:

$$\begin{cases} 2x+y=-1 \\ -3x+y=-2 \\ -x+y=1 \end{cases} \qquad \begin{cases} 2x+y=1 \\ -2x+y=1 \\ -x+y=1 \end{cases} \qquad \begin{cases} x-y=0 \\ 2x-2y=-2 \\ -x+y=1 \end{cases} \qquad \begin{cases} 2x+y=1 \\ 2x-2y=-2 \\ -x+y=1 \end{cases}$$

How many systems above have solutions?

a) 0

b) 1

c) 2

d) 3

e) 4

6. [objective function] How many of the following four objective functions are linear in terms of their parameters?

$$f_1(x) = w_1 + w_2 x + w_3 x^2$$
$$f_2(x) = w_1 + w_2 e^x + w_3 e^{x^2}$$
$$f_3(x) = w_1 + e^{w_2 x} + e^{w_3 x^2}$$
$$f_4(x) = w_1 + \log(x^{w_2}) + \log(x^{2w_2})$$

a) 0

b) 1

c) 2

d) 3 Lab 4

e) 4

7. [gradient descent] Which one of the following is incorrectly described?

Lab 2 Slide 2

a) A stochastic gradient descent (SGD) algorithm must update parameters after calculating the gradient of the objective functions for all data points in a random order.

b) A stochastic gradient descent (SGD) algorithm can update parameters after calculating the gradient of the objective functions for a subset of all data points.

c) A stochastic gradient descent (SGD) algorithm can update parameters after calculating the gradient of the objective function for a single data point.

d) A standard gradient descent (GD) method can update parameters after calculating the gradient of the objective functions for all data points.

e) A standard gradient descent (GD) method can update parameters after calculating the gradient of the objective functions for all data points in a random order.

Stochastic Gradient Descent Algorithm

If the number of data points n is small, gradient descent is fine, but sometimes (eg. 'Big Data') n could be a billion.

- stochastic gradient descent is more similar to perceptron.
- it looks at gradient of one data point at a time rather than summing across *all* data points.
- this gives a stochastic estimate of gradient.

8. [likelihood] The likelihood of a parameter set θ for observed data x is

Likelihood

The likelihood of parameter values θ given data $\mathbf{x}$ is the probability for those observed data given those parameter values:

$$\mathcal{L}(\theta|x) = p(x|\theta)$$

a) $p(x|\theta)p(\theta)$

Slide 3 b) $p(x|\theta)$

Slide 6 c) $\dfrac{p(x|\theta)p(\theta)}{p(x)}$

d) $p(\theta|x)$

e) $p(\theta|x)p(x)$

Posterior Distribution

Posterior distribution is found by combining the prior with the likelihood.

- posterior distribution to represent our belief *after* we see the data of the likely value of the parameters
- the posterior being derived through **Bayes' rule**:

$$p(c|y) = \frac{p(y|c)p(c)}{p(y)}$$

9. [likelihood] Which one of the following descriptions about a log likelihood is correct?

Slide 3 a) The logarithm is a monotonically increasing decreasing function, hence the log of a likelihood achieves its maximum at the point where the likelihood achieves its maximum minimum.

Slide 6 b) A posterior pdf (probability density function) is calculated as a product of the log likelihood and the prior pdf. $p(c|\mathbf{x}, \mathbf{y}, m, \sigma^2) = \dfrac{p(\mathbf{y}|\mathbf{x}, m, c, \sigma^2)p(c)}{p(\mathbf{y}|\mathbf{x}, m, \sigma^2)}$ (PDF)

c) A number of well known distributions, such as the normal distribution, have likelihood functions that contain products of factors involving exponentiation. The logarithm of such a function repllaces products with sums, hence easier to differentiate than the original function.

Slide 3 d) Probabilistic interpretation for an error function is a log likelihood of normal distribution. Thus maximising error function is equivalent to maximum likelihood with respect to parameters.

Error Function

Probabilistic interpretation for minimising error function is equivalent to maximum likelihood with respect to parameters.

$$\mathcal{L}(m, c, \sigma^2) = \underbrace{-\frac{n}{2}\log 2\pi}_{\text{constant}} - E(m, c, \sigma^2)$$

- minimising error function is equivalent to maximising log likelihood
- maximising log likelihood is equivalent to maximising the likelihood because the logarithm is monotonic function

Slide 3 e) The log likelihood of the normal distribution involves exponentiation.

Log Likelihood

Often simpler if we work with the log likelihood:

$$\mathcal{L}(m, c, \sigma^2) \triangleq \log p(\mathbf{y}|\mathbf{x}, m, c, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2}$$

10. [generalisation] Which of the following describes cross validation incorrectly?

Lab 5 a) It assesses how the results of a statistical analysis will generalise to an unseen data set.

Slide 5 b) It tests the model using the validation dataset in order to reduce the chance of an overfitting problem.

c) It partitions data into complementary subsets for training and for validation, which is repeated using different partitions and finally the validation results are averaged.

d) Leave-one-out validation is its special case, where partition is made with one validation data point and the rest for training data.

e) It produces meaningful results if the validation set and training set are drawn from the totally different population.

11. [unsupervised learning] Which one of the following is not considered as an approach to unsupervised learning?

Slide 7

Supervised Learning

Supervised learning is learning where each data has a label.

- training example: set of pairs consisting of an input object and a desired output value
- recall the bias-variance tradeoff (**week 5**)

(eg) regression analysis

a) support vector machine (SVM)

b) k-means clustering

c) latent variable models

d) factor analysis

e) principal component analysis (PCA)

Unsupervised Learning

In unsupervised learning we have no labels for the data. It is often thought of as structure discovery, such as finding features in the data.

Unsupervised learning techniques:

- clustering (eg: k-means)
- latent variable models (eg: EM algorithm)
- blind signal separation (eg: PCA, ICA, SVD)

Questions 12 to 15 are about python code. The `numpy` library has been imported as `np` and we are given `u` and `v` as one dimensional numpy arrays. Note that `numpy.outer` calculates the outer product of two vectors.

12. [python code] Which one of the following pieces of python code calculates $\mathbf{Y} = \mathbf{u}\mathbf{v}^\top$? Note that $\mathbf{u}$ and $\mathbf{v}$ are both column vectors.

a) `y = u*v`

b) `y = np.sum(u*v)`

c) `y = np.sum(v*u**2)`

d) `y = np.sum(u*v)**2`

e) `y = np.outer(u,v)`

13. [python code] Which one of the following pieces of python code calculates $y = \mathbf{v}^\top\mathbf{u}\mathbf{u}^\top\mathbf{v}$? Note that $\mathbf{u}$ and $\mathbf{v}$ are both column vectors.

a) `y = u*v`

b) `y = np.sum(u*v)`

c) `y = np.sum(v*u**2)`

d) `y = np.sum(u*v)**2`

e) `y = np.outer(u,v)`

14. [python code] Which one of the following pieces of python code calculates $y = tr(\mathbf{u}\mathbf{v}^\top)$? Note that $\mathbf{u}$ and $\mathbf{v}$ are both column vectors, and that $tr(\mathbf{A})$ indicates the sum of the diagonal elements of the square matrix $\mathbf{A}$.

a) `y = u*v`

b) `y = np.sum(u*v)`

c) `y = np.sum(v*u**2)`

d) `y = np.sum(u*v)**2`

e) `y = np.outer(u,v)`

15. [python code] Which one of the following pieces of python code calculates $y = \mathbf{v}^\top diag(\mathbf{u})$? Note that $\mathbf{u}$ and $\mathbf{v}$ are both column vectors, and that $diag(\mathbf{z})$ forms a diagonal matrix with diagonal elements given by elements of $\mathbf{z}$.

a) `y = u*v` 位乘

b) `y = np.sum(u*v)`

c) `y = np.sum(v*u**2)`

d) `y = np.sum(u*v)**2`

e) `y = np.outer(u,v)`