ML Residency Test

nttuan8

October 2021

1. We have seen some data D which we try to represent using parameter θ. Which combination of "semantic" names for the distribution below is correct?

$$\underbrace{p(\theta|D)}_{\text{A}} = \frac{\overbrace{p(D|\theta)}^{\text{B}}\overbrace{p(\theta)}^{\text{C}}}{\underbrace{p(D)}_{\text{D}}} \tag{1}$$

   (a) {A, B, C, D} = {Likelihood, Prior, Evidence, Posterior}
   (b) {A, B, C, D} = {Posterior, Likelihood, Prior, Evidence}
   (c) {A, B, C, D} = {Posterior, Joint, Prior, Evidence}
   (d) {A, B, C, D} = {Evidence, Posterior, Likelihood, Prior}

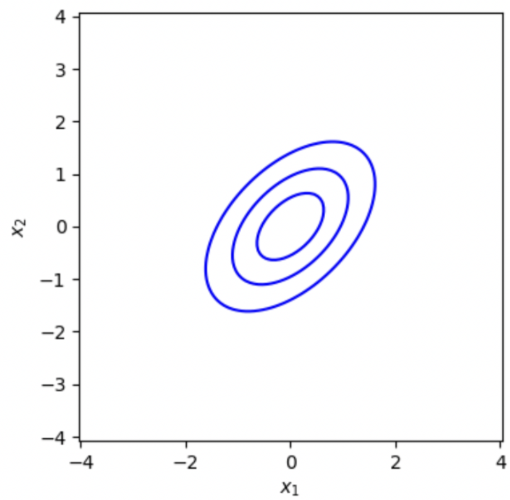
2. As the number of data points grows approaching infinity will a Maximum-a-posterior always be the same as the Maximum Likelihood estimate?

   (a) Yes
   (b) No

3. To find the minima of a continous non-explicit function, we can use Bayesian optimization. Which of the following statements is true for Bayesian optimization?

   (a) We are guaranteed to find the global minima of the funciton in finite time
   (b) The number of tests it takes for us to reach a solution is independent of our prior assumption of the funciton that we minimize.
   (c) It is impossible to know how close to the true minima our current estimate is
   (d) None of the above

4. Match the Gaussian in the image below to the correct co-variance matrix.



(a) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

(b) $\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$

(d) $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$

5. Given a set of associated input values X and target values t you derived the posterior distribution over regression weights w for a linear model

$$p(w|t, X, \alpha, \beta)$$

where α and β are parameters of the likelihood and the prior respectively. In order to reach the predictive distribution of the model which random variable should be marginalize over?

(a) X the input values of the training data

(b) α and β the parameters of the likelihood and the prior

(c) w the regression weights

(d) t the target values of training data

6. What strategies can help reduce overfitting in decision trees?

(a) Pruning

(b) Enforce a minimum number of samples in leaf node

(c) Make sure each leaf node is one pure class

(d) Enforce a maximum depth for the tree

7. How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression?

   (a) Ridge has larger bias, larger variance

   (b) Ridge has smaller bias, larger variance

   (c) Ridge has larger bias, smaller variance

   (d) Ridge has smaller bias, smaller variance

   c

8. Which of the following are true about bagging?

   (a) In bagging, we choose random subsamples of the input points with replacement

   (b) The main purpose of bagging is to decrease the bias of learning algorithm

   (c) Bagging is ineffective with logistic regression, because all of the learners learn exactly the same decision boudary

   (d) If we use decision trees have one sample point per leaf, bagging never gives lower training error than one ordinary decision tree

9. Suppose your model í overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?

   (a) Increase the amount of training data

   (b) Improve the optimization algorithm being used for error minimization

   (c) Decrease the model complexity

   (d) Reduce the noise in the training data

10. Which of the followings are used to access a classification model?

    (a) Confusion matrix

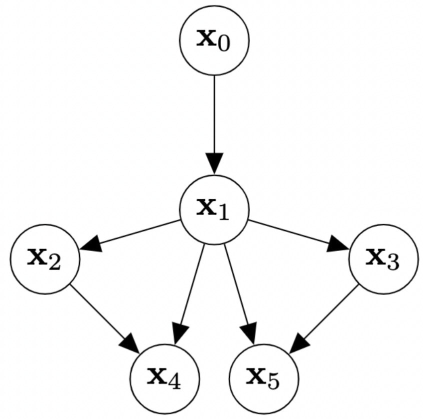
    (b) Mean absolute error

    (c) Area under the ROC curve

    (d) All of the above

11. A Graphical model is a visual description of the joint distribution factorised into its components. Which factorisation does the following model encode?

    (a) $p(x_4, x_5|x_1, x_2, x_3)p(x_2)p(x_3)p(x_1|x_0)$



(b) $p(x_4, x_5|x_1, x_2, x_3)p(x_1|x_0)$

(c) $p(x_5|x_1, x_3)p(x_4|x_1, x_2)p(x_2|x_1)p(x_3|x_1)p(x_1|x_0)p(x_0)$

(d) $p(x_5|x_1, x_2, x_3)p(x_4|x_1, x_2, x_3)p(x_2)p(x_3)p(x_1|x_0)p(x_0)$

12. Which of the following is a reasonable way to select the number of principal components "k"?

    (a) Choose k to be the smallest value so that at least 99% of the variance is retained

    (b) Choose k to be the largest value so that at least 99% of the variance is retained

    (c) Choose k to be 99% of m (m is the dimension of input)

    (d) Use elbow method

13. How do you handle missing or corrupted data in a dataset?

    (a) Drop missing rows or columns

    (b) Assign a unique category to missing values

    (c) Replace missing values with mean/median/mode

    (d) All of the above

14. The Laplace approximation is a method to approximate an intractable posterior distribution. Which of the following statements is true for the Laplace approximation?

    (a) The Laplace approximation is exact

    (b) We need to be able to find the maximum of the posterior to apply the approximation

(c) The optimization problem the approximation leads to is non-convex

(d) None of the above

15. Which of the following statements are true for sampling?

    (a) Using sampling we try to approximate an intractable integral with a sum

    (b) The more dependent the samples we use in the approximation the less samples we are likely to need

    (c) A sampling method is an example of a deterministic approximation and will recover the same solution every time it is applied

    (d) None of the above