

Introduction to Machine Learning, Spring 2023

Homework 2

(Due Thurs, Mar. 23 at 11:59pm (CST))

March 14, 2023

1. [15 points] Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $\mathbf{x} \in \mathbb{R}^d$ to high dimensional space Q by giving the form of dot product in Q : $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

- (a) Prove that the kernel is symmetric, i.e. $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$. [5 points]

solution:

- (b) Assume we use radial basis kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$. Thus there is some implicit unknown mapping function $\phi(\mathbf{x})$. Prove that for any two input instances \mathbf{x}_i and \mathbf{x}_j , the squared Euclidean distance of their corresponding points in the feature space Q is less than 2, i.e. prove that $\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 \leq 2$. [5 points]

solution:

- (c) With the help of a kernel function, SVM attempts to construct a hyper-plane in the feature space Q that maximizes the margin between two classes. The classification decision of any \mathbf{x} is made on the basis of the sign of

$$\langle \hat{\mathbf{w}}, \phi(\mathbf{x}) \rangle + \hat{w}_0 = \sum_{i \in SV} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + \hat{w}_0 = f(\mathbf{x}; \alpha, \hat{w}_0),$$

where $\hat{\mathbf{w}}$ and \hat{w}_0 are parameters for the classification hyper-plane in the feature space Q , SV is the set of support vectors, and α_i is the coefficient for the i -th support vector. Again we use the radial basis kernel function. Assume that the training instances are linearly separable in the feature space Q , and assume that the SVM finds a margin that perfectly separates the points.

If we choose a test point \mathbf{x}_{far} which is far away from any training instance \mathbf{x}_i (distance here is measured in the original space \mathbb{R}^d), prove that $f(\mathbf{x}_{far}; \alpha, \hat{w}_0) \approx \hat{w}_0$. [5 points]

solution:

2. [15 points] The Poisson distribution is a useful discrete distribution which can be used to model the number of occurrences of something per unit time. For example, in networking, the number of packets to arrive in a given time window is often assumed to follow a Poisson distribution. If X is Poisson distributed, i.e. $X \sim \text{Poisson}(\lambda)$, 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 i.i.d. data points from Poisson (λ) : $\mathcal{D} = \{X_1, \dots, X_n\}$ (For the purpose of this problem, you can only use the knowledge about the Poisson and Gamma distributions provided in this problem.)

- (a) Show that the sample mean $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i$ is the maximum likelihood estimate (MLE) of λ and it is unbiased ($\mathbb{E}(\hat{\lambda}) = \lambda$). [5 points]

solution:

- (b) Now let's be Bayesian and put a prior distribution over λ . Assuming that λ follows a Gamma distribution with the parameters (α, β) , its probability density function:

$$p(\lambda | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

where $\Gamma(\alpha) = (\alpha - 1) !$ (here we assume α is a positive integer). Compute the posterior distribution over λ .

[5 points]

solution:

- (c) Derive an analytic expression for the maximum a posterior (MAP) of λ under Gamma (α, β) prior. [5 points]

solution:

3. [10 points] using d-separation on figure?? to discuss the following questions.

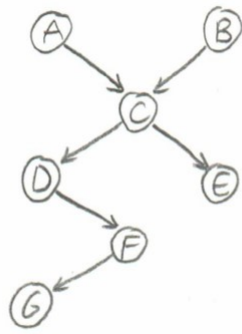


Figure 1: A Bayes net

(a) Are A and B conditionally independent, given D and F? [5 points]

solution:

(b) $P(D|CEG) = ?P(D|C)$ [5 points]

solution: