Questions

1. **Karataka** gave the following regression dataset to **Damanaka**: $\mathcal{D} = \{(0.1, -10), (\pi, 20), (\sqrt{3}, 5)\}$, where the first entry in each pair is the input and the second is the output. Further, **Karataka** instructed **Damanaka** to employ the generative KDE regression set-up. After training, **Damanaka** claims that the predicted label with the trained model at $x = 9$ is $y = 25$. **Damanaka**'s claim is definately false

   [[*Fill this blank with either "is definitely false" or "may be true" or "cannot be validated from the given information", while providing justification in the box below* **1.5 Marks**]]:

   $\hat{f}(x) = \sum_{i=1}^{m} \rho_i y_i$ where $\rho_i \geq 0$, $\sum_i \rho_i = 1$. Hence $\hat{f}(x)$ is an interpolation of $-10, 5, 20$. But $25 \notin [-10, 20]$

   extrapolation

   interpolation

2. Consider the following stochastic optimization problem:

$$\min_{x \in \mathbb{R}^n} \mathbb{E}[f(x, W)], \tag{1}$$

   where W is a random variable having some fixed, yet unknown, distribution and $f : \mathbb{R}^n \times \mathbb{R}^d \mapsto \mathbb{R}$ is a given function. Let $\begin{bmatrix} g_1(x, w) \\ g_2(x, w) \end{bmatrix}$ denote the gradient of f at (x, w), where $g_1 : \mathbb{R}^n \times \mathbb{R}^d \mapsto \mathbb{R}^n$. Let the update step at the k^{th} iteration in the general SGD algorithm for solving (1) be $x^{(k+1)} \equiv x^{(k)} - \eta r_k$, where r_k is an instantiation of a random direction R_k. Let $X^{(k)}$ denote the random variable denoting the (random) iterate of the variable at the k^{th} iteration i.e., $x^{(k)}$ is an instantiation of $X^{(k)}$. Then, the only technical condition on the random direction R_k is given by the expression:

   $E\left[R_k \middle| X^{(k)} = x^{(k)}\right] = E\left[g_1(x^{(k)}, W)\right]$.

   [[*Fill blank with an expression only involving one or more of the following (repetition allowed): (i) common math symbols like equality/set-belongs-to, addition, expectation/conditional-expectation etc., (ii) k, (iii) r, (iv) R, (v) X (vi) x (vii)* g_1 *(viii)* g_2, *(ix)* W, *(x)* w. **1 Mark**]] If samples for W are $w_1, \ldots, w_m$, then, few specific ways for defining the random direction R_k were taught to you in the lecture. Instantiations of two different R_k are: $r_k = g_1(x^{(k)}, w_k)$, and

   $r_k = 0.5\, g_1(x^{(k)}, w_{2k-1}) + 0.5\, g_1(x^{(k)}, w_{2k})$

   Any other minibatch style answer is also fine.

   [[*Again, for filling these two blanks, you are allowed to use only the symbols allowed for the previous blank. Additionally, you may use some or all of* $w_1, \ldots, w_m$. **0.5Mark+1 Mark.**]]

3. Consider a regression problem where input space, $\mathcal{X} = \mathbb{R}$, and output space, $\mathcal{Y} = \mathbb{R}$, and the model set-up is the generative linear regression taught in lecture. Let $p^*(x, y)$ denote the underlying concept relating the inputs and labels and f^* be the corresponding Bayes optimal. The object that is modelled in this set-up is $p^*(x, y)$ [[*Fill in this blank with either* $p^*(x, y)$ *or* $p^*(y/x)$ *or* f^*. **(0.25Mark)**]]. This object is modelled using the Gaussian model [[*Fill in this blank with a proper noun, which is the name of one of the models taught in this course.* **(0.25Mark)**]]. The stochastic optimization problem defining the "best" parameter of this model is given by:

   Multivariate Gaussian or exponential family model are also correct.

$$\underset{\substack{\mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R} \\ \Sigma \succ 0}}{\text{argmin}}\; E_{(X,Y) \sim p^*}\left[[X - \mu_1, \; Y - \mu_2]\, \Sigma^{-1} \begin{bmatrix} X - \mu_1 \\ Y - \mu_1 \end{bmatrix} \right] + \log|\Sigma|$$

[[*Your expression must be a simplified expression in terms of the (specific) parameters of the model. No marks will be given if a general stochastic optimization problem is written or if written without simplification.* **3/4 Mark**]]. Let the training dataset be $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. While the above stochastic optimization problem can be solved using SGD, the optimization problem corresponding to the Stochastic Average Approximation (SAA) can be written as:

$$\underset{\substack{\mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R} \\ \Sigma \succ 0}}{\text{argmin}} \frac{1}{m} \sum_{i=1}^{m} [x_i - \mu_1 \quad y_i - \mu_2] \Sigma^{-1} \begin{bmatrix} x_i - \mu_1 \\ y_i - \mu_2 \end{bmatrix} + \log |\Sigma|$$

[[*Your expression must be a simplified expression in terms of the (specific) parameters of the model and elements of $\mathcal{D}$.* **3/4 Mark**]]. Now, say the training set is actually $\mathcal{D} = \{(3, 3), (2, 0), (-1, -1), (0, 2)\}$. Then, as per Murphy's book, the optimal solution for the above optimization problem is given by:

$$\hat{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{bmatrix}$$

[[*Fill the blank with equation(s) where LHS denotes the parameter and RHS is the optimal estimate of it, written in terms of decimal numbers.* **1 Mark**]]. The Bayes optimal corresponding to the above parameter estimate is given by:

$$\hat{f}(x) = 0.6\, x + 0.4.$$

[[*Fill these two blanks with decimal numbers.* **1 Mark**]]

4. Consider the 3-class classification training dataset:

$$\mathcal{D} = \{(-1, \Xi), (-0.5, \Xi), (0, \ddagger), (1, ¶), (2, ¶), (3, ¶)\}.$$

Let's employ the Bayes classifier setup with tied variance, σ^2. Let μ_1, μ_2, μ_3 denote the means of the class conditionals of $\Xi, \ddagger, ¶$ respectively. Recall that the MLE problem always decouples into separate optimization problems wrt. the parameters of the class conditionals and the parameters of the label prior. For this dataset, the MLE based label prior estimate is $\hat{p}(\Xi) = 1/3$, $\hat{p}(\ddagger) = 1/6$, $\hat{p}(¶) = 1/2$ [[*Fill these three blanks with numbers using fractional notation.* **0.5 Mark**]]. For this dataset, the MLE optimization problem wrt to the parameters $(\mu_1, \mu_2, \mu_3, \sigma^2)$ is:

$$\arg \max_{\mu_1, \mu_2, \mu_3, \sigma > 0} -\frac{1}{\sigma^2}\left((-1-\mu_1)^2 + (-0.5-\mu_1)^2 + \mu_2^2 + (1-\mu_3)^2 + (2-\mu_3)^2 + (3-\mu_3)^2\right) + 6 \log 1/\sigma^2$$

[[*Your expression in the above blank must be specific to the given training data and simplified.* **1 Mark**]]. Solving for (i.e., eliminating) μ_1, μ_2, μ_3 first gives the estimates for them as $\hat{\mu}_1 = -0.75$, $\hat{\mu}_2 = 0$, $\hat{\mu}_3 = 2$. Then, solving for σ^2 gives estimate for it as $\hat{\sigma}^2 = 17/48$ [[*Fill these four blanks using decimal numbers. You can also use fractional notation for numbers.* **(0.5Mark+1Mark)**]]. The corresponding prediction function is given by $\hat{f}(x) = \begin{cases} \Xi & \text{if } x \leq g(\hat{\mu}, \hat{\sigma}^2, \hat{p}) \\ ¶ & \text{if } x \geq h(\hat{\mu}, \hat{\sigma}^2, \hat{p}) \\ \ddagger & \text{otherwise.} \end{cases}$ where

$$g(\hat{\mu}, \hat{\sigma}^2, \hat{p}) = \min\left(\frac{\hat{\mu}_2 + \hat{\mu}_1}{2} + \hat{\sigma}^2 \frac{\log \hat{\theta}_1/\hat{\theta}_2}{\hat{\mu}_2 - \hat{\mu}_1}, \frac{\hat{\mu}_3 + \hat{\mu}_1}{2} + \hat{\sigma}^2 \frac{\log \hat{\theta}_1/\hat{\theta}_3}{\hat{\mu}_3 - \hat{\mu}_1}\right),$$

$$h(\hat{\mu}, \hat{\sigma}^2, \hat{p}) = \max\left(\frac{\hat{\mu}_1 + \hat{\mu}_3}{2} + \hat{\sigma}^2 \frac{\log \hat{\theta}_3/\hat{\theta}_1}{\hat{\mu}_1 - \hat{\mu}_3}, \frac{\hat{\mu}_2 + \hat{\mu}_3}{2} + \hat{\sigma}^2 \frac{\log \hat{\theta}_3/\hat{\theta}_2}{\hat{\mu}_2 - \hat{\mu}_3}\right).$$

[[*Fill the above blanks with expressions involving the estimated prior and class conditional parameters.* **1 Mark**]]

5. Consider the c-class logistic regression set-up taught in lecture. Here, the expression for the model likelihood, $p(y/x)$, in terms of the parameters and the feature map is: $e^{w_y^T\phi(x)} / \sum_{i=1}^{c} e^{w_i^T\phi(x)}$.

[1 Mark]

6. The key assumption that is unique in nearest neighbour classifier set-up, which is foundational in the theorem proving it's asymptotic correctness is:

$$\boxed{|p(y/x) - p(y/x')| \leq L\, d(x, x')}$$

→ Any informal equivalent is also fine.

[0.5 Mark]

7. With respect to k-NN models, statement(s) B, C, D, E, among the ones below, is(are) false, and the remaining are(is) true:

   A. k-NN models are non-parametric.

   B. 1-NN classifier is guaranteed to achieve 0 generalization error, provided $m \to \infty$.

   C. 1-NN classifier is guaranteed to achieve 0 estimation error, provided $m \to \infty$.

   D. k-NN models may suffer from the curse of dimensionalty, but are computationally attractive for high-dimensional data.

   E. k-NN models do not suffer from the curse of dimensionalty, but are computationally in-attractive for high-dimensional data.

   [[*Fill the above blank with one or more of "A", "B", "C", "D", "E".* **1 Mark**]].

8. Recall that a smoothing kernel needs to satisfy two technical conditions. Now, let $\kappa_1 : \mathcal{X} \mapsto \mathbb{R}_+, \kappa_2 : \mathcal{X} \mapsto \mathbb{R}_+, \kappa_3 : \mathcal{Y} \mapsto \mathbb{R}_+$ be three valid smoothing kernels. Then, the statement "κ_4 defined by $\kappa_4(x) \equiv \kappa_1(x)\kappa_2(x)$ is always a valid smoothing kernel" is FALSE. The statement "κ_5 defined by $\kappa_5(x, y) \equiv \kappa_1(x)\kappa_3(y)$ is always a valid smoothing kernel" is TRUE [[*Fill in these two blanks with either "TRUE" or "FALSE". Justify your answer in the box below for the blanks you filled with "TRUE"*]]:

$k_1(x) \geq 0, k_3(y) \geq 0 \Rightarrow k_5(x,y) \geq 0$. $\int k_5(x,y)\,dx\,dy = \int k_1(x)\,dx \int k_3(y)\,dy = 1$	$k_5(-x,-y) = k_1(-x)\,k_3(-y) = k_1(x)\,k_3(y) = k_5(x,y)$ etc.

[0.5 Mark]

① $\hat{f}(x) = \sum_{i=1}^{m} \rho_i y_i = \rho_1(-10) + \rho_2(20) + \rho_3(5) \in [-10, 20]$.

interpolation

$\because \rho_i \geq 0, \ \sum_i \rho_i = 1$

Bail 25 £

② $E[R_k \mid X^{(k)} = x^{(k)}] = \nabla_x E[f(x, W)]\Big|_{x = x^{(k)}}$

$= E\left[\nabla_x f(x, W)\Big|_{x = x^{(k)}}\right]$

$= E\left[g_1(x^{(k)}, W)\right]$

$g_1(x^{(k)}, w_k)$ is sample version of

is also a sample version

Similarly, using <u>minibatch</u> ideas, $0.5\, g_1(x^{(k)}, w_{2k-1}) + 0.5\, g_1(x^{(k)}, w_{2k})$

③ The general form of the stochastic optimisation for likelihood estimation is

$$\underset{q \in Q}{\arg\max}\ E_{X \sim p^*}\left[\log q(X)\right]$$

(equation 2.11 in summary notes)

In our case, $q(x) \Rightarrow$ Gaussian in $\mathbb{R}^2 \rightarrow p(x,y) \propto \dfrac{e^{-\frac{1}{2}[x-\mu_1,\ y-\mu_2]\,\Sigma^{-1}\begin{bmatrix} x-\mu_1 \\ y-\mu_2 \end{bmatrix}}}{|\Sigma|^{1/2}}$

$$\underset{\substack{\mu_1 \in \mathbb{R},\, \mu_2 \in \mathbb{R} \\ \Sigma \succ 0 \in \mathbb{R}^2}}{\arg\cancel{\max}\ \min}\ E_{(X,Y) \sim p^*}\left[[X-\mu_1,\ Y-\mu_2]\,\Sigma^{-1}\begin{bmatrix} X-\mu_1 \\ Y-\mu_2 \end{bmatrix}\right] + \log|\Sigma|$$

$\mu \longrightarrow$ MLE estimate is sample mean $= \begin{bmatrix} \frac{1}{4}(3+2+-1+0) \\ \frac{1}{4}(3+0-1+2) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\Sigma \longrightarrow$ MLE estimate is sample covariance $= \frac{1}{4}\left[\begin{bmatrix} 2 \\ 2 \end{bmatrix}\begin{bmatrix} 2 & 2 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix} + \begin{bmatrix} -2 \\ -2 \end{bmatrix}\begin{bmatrix} -2 & -2 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \end{bmatrix} \right]$

$$= \begin{bmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{bmatrix}$$

$$\hat{f}(x) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x-\mu_1)$$ (Equation in sect 5.2 in summary notes)

$$= 1 + \frac{3}{2}\,\frac{2}{5}(x-1) = 0.6x + 0.4$$

~~$\hat{p}(\Xi) = \frac{2}{6} = \frac{1}{3}$ $\hat{p}(\ddagger) = \frac{1}{6}$ $\hat{p}(9) = \frac{3}{6} = \frac{1}{2}$~~

~~Let~~ ~~Let~~ ~~argmax $\mu_1, \mu_2, \mu_3, \sigma, \hat{p}\theta$~~

4 In Bayes classifier, $\hat{p}(x,y) = \hat{p}(x/y)\,\hat{p}(y)$ is modelled.

~~Let~~ μ_y, Σ_y ~~be parameters~~

Let parameters for class conditionals

$\hat{p}(x/\Xi) \approx N(\mu_1, \sigma^2)$

$\hat{p}(x/\ddagger) \approx N(\mu_2, \sigma^2)$

$\hat{p}(x/9) \approx N(\mu_3, \sigma^2)$

Let parameters be $\theta_1, \theta_2, \theta_3$.

$\hat{p}(\Xi)$ $\hat{p}(\ddagger)$ $\hat{p}(9)$

MLE is:

$$\underset{\substack{\mu_1,\mu_2,\mu_3\in\mathbb{R}\\ \sigma>0\\ \theta_1,\theta_2,\theta_3\geq 0\\ \sum_i \theta_i=1}}{\operatorname{argmax}} \sum_{i=1}^{m} \log\left(p(x_i/y_i)\,p(y_i)\right)$$

$$= \underset{\substack{\mu_1,\mu_2,\mu_3\in\mathbb{R}\\ \sigma>0}}{\operatorname{argmax}} \sum_{i=1}^{m} \log p(x_i/y_i) + \boxed{\underset{\substack{\theta_1,\theta_2,\theta_3\geq 0\\ \sum_i\theta_i=1}}{\operatorname{argmax}} \sum_{i=1}^{m} p(y_i)}$$

$\to$ Multinoulli MLE

$\therefore \hat\theta_1 = 2/6 = 1/3 = \hat p(\ominus)$

$\hat\theta_2 = 1/6 = \hat p(\mp)$

$\hat\theta_3 = 3/6 = \hat p(\$)$

$= 1/2$

$$= \underset{\substack{\mu_1,\mu_2,\mu_3\in\mathbb{R}\\ \sigma>0}}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\Big(\underbrace{(-1-\mu_1)^2+(-0.5-\mu_1)^2}_{\text{Solving}}+\underbrace{(\mu_2)^2}_{\text{Solving}}+\underbrace{(1-\mu_3)^2+(2-\mu_3)^2+(3-\mu_3)^2}_{\text{Solving}}\Big) + \frac{6}{2}\log 1/\sigma^2$$

$\hat\mu_1 = \frac{-1+-0.5}{2} = -0.75$

$\hat\mu_2 = 0$

$\hat\mu_3 = \frac{1+2+3}{3} = 2$

$\frac{1}{\sigma^2} = x$

$$= \underset{x>0}{\operatorname{argmin}}\ \frac{x}{2}\left((-1+0.75)^2+(-0.5+0.75)^2+(1-2)^2+(3-2)^2\right) - 6\log x$$

$$= \underset{x>0}{\operatorname{argmin}}\ \underbrace{\frac{17x}{8} - 6\log x}_{g(x)} = \frac{48}{17} \Rightarrow \hat\sigma^2 = \frac{1}{x} = \frac{17}{48}$$

$g'(x) = +\frac{17}{8} - \frac{6}{x} = 0$

Let $(\mathrm{I}) \equiv \log\left(\hat{p}(x/\ominus)\,\hat{p}(\ominus)\right)$

$$= -\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_1)^2 - \frac{1}{2}\log\hat{\sigma}^2 + \log\hat{\theta}_1$$

$(\mathrm{II}) \equiv \log\left(\hat{p}(x/\ddagger)\,\hat{p}(\ddagger)\right)$

$$= -\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_2)^2 - \frac{1}{2}\log\hat{\sigma}^2 + \log\hat{\theta}_2$$

$(\mathrm{III}) \equiv \log\left(\hat{p}(x/g)\,\hat{p}(g)\right)$

$$= -\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_3)^2 - \frac{1}{2}\log\hat{\sigma}^2 + \log\hat{\theta}_3$$

$\hat{f}(x) = \ominus$ iff $(\mathrm{I}) \geq (\mathrm{II}),\ (\mathrm{I}) \geq (\mathrm{III})$

i.e. iff $-(x-\hat{\mu}_1)^2 + 2\hat{\sigma}^2\log\hat{\theta}_1 \geq -(x-\hat{\mu}_2)^2 + 2\hat{\sigma}^2\log\hat{\theta}_2$

$$\Leftrightarrow\ x \geq \frac{\hat{\sigma}^2(\log\hat{\theta}_1/\hat{\theta}_2)}{\hat{\mu}_2-\hat{\mu}_1} + \frac{\hat{\mu}_2+\hat{\mu}_1}{2}$$

|||ly $(\mathrm{I}) \geq (\mathrm{III}) \Leftrightarrow x \geq \dfrac{\hat{\sigma}^2\log(\hat{\theta}_1/\hat{\theta}_3)}{\hat{\mu}_3-\hat{\mu}_1} + \dfrac{\hat{\mu}_1+\hat{\mu}_3}{2}$

$$\therefore\ \hat{f}(x) = \ominus \text{ iff } x \geq \max\left(\frac{\hat{\sigma}^2\log(\hat{\theta}_1/\hat{\theta}_2)}{\hat{\mu}_2-\hat{\mu}_1} + \frac{\hat{\mu}_2+\hat{\mu}_1}{2},\ \frac{\hat{\sigma}^2\log(\hat{\theta}_1/\hat{\theta}_3)}{\hat{\mu}_3-\hat{\mu}_1} + \frac{\hat{\mu}_1+\hat{\mu}_3}{2}\right)$$

$\hat{f}(x) = g$ iff $(\mathrm{III}) \geq (\mathrm{I}),\ (\mathrm{III}) \geq (\mathrm{II})$ guess by symmetry

$$\Leftrightarrow\ x < \frac{\hat{\sigma}^2\log(\hat{\theta}_3/\hat{\theta}_1)}{\hat{\mu}_1-\hat{\mu}_3} + \frac{\hat{\mu}_3+\hat{\mu}_1}{2}, \quad x < \frac{\hat{\sigma}^2\log(\hat{\theta}_3/\hat{\theta}_2)}{\hat{\mu}_2-\hat{\mu}_3} + \frac{\hat{\mu}_2+\hat{\mu}_3}{2}$$