1

(a) You are given a dataset $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ where $\boldsymbol{x}_n \in \mathbb{R}^p$ is a p-dimensional vector of real-valued features associated with an individual n and y_n is a real-valued “output" variable for n. Show how you would set up a **linear regression** model by expressing y_n as a **linear** function of $\boldsymbol{x}_n$. What would the learning algorithm return? *[5 marks]*

(b) For a new data ... N, with scalars $x_n, y_n \in \mathbb{R}$ ... x_n, how would the correspond ... $\boldsymbol{A}$ ... $N = 5$ ata points describe the entries of the rows and columns of ... unction. *[8 marks]*

(c) You are asked to make an update to the weights of the quadratic function $y = f(x) = w_0 + w_1 x + w_2 x^2$ from a given set of values $\boldsymbol{w} = (w_0, w_1, w_2) = (2, 1, -2)$ to minimise the squared loss on a minibatch consisting of the two points:

$$\{(x_1, y_1), (x_2, y_2)\} = \{(1.0, 0.5), (-1.0, -0.5)\}.$$

In stochastic gradient descent, what would the direction of the gradient be for each point? What would the batch gradient be? For a learning rate of 0.1 what is the updated weight vector $\boldsymbol{w}$ when the entire minibatch is used to compute th *[12 marks]*

2

(a) For a data set $\{\boldsymbol{x}_n | n = 1, \ldots, N\}$ of N column vectors $\boldsymbol{x}_n$ with feature values as components $x_{n,1}, x_{n,2}, \ldots, x_{n,p}$, describe in detail how you would construct the covariance matrix of features. *[6 marks]*

(b) Using the singular value decomposition (SVD) of a $(N \times p)$ matrix $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ describe how you would construct a rank k approximation of $\boldsymbol{A}$, for $k < \min\{N, p\}$. *[6 marks]*

(c) Explain in detail how you would use Principal Components Analysis (PCA) to reduce the dimensionality p of the vectors $\boldsymbol{x}_n$: $\mathbb{R}^p \ni \boldsymbol{x}_n \mapsto \boldsymbol{x}'_n \in \mathbb{R}^k$, $k < p$. Comment on the relationship between the variance of the reduced dimensional version of the vectors $\boldsymbol{x}'_n$ and those of the original $\boldsymbol{x}_n$. *[7 marks]*

(d) For classification problems where there is a training set of labelled data, explain why dimensionality reduction using PCA may give rise to poor classification outcomes. You may draw a figure to illustrate your arguments. *[6 marks]*

3

(a) Describe the k-means clustering algorithm. *[7 marks]*

(b) In Fisher's Linear Discriminant Analysis (LDA), training data – $\boldsymbol{x}_A^i$ and $\boldsymbol{x}_B^j$ from 2 classes A and B – is projected along some vector $\boldsymbol{w}$ so that $\boldsymbol{x}_A^i$, $\boldsymbol{x}_B^j$ are mapped into

$$y_A^i := \boldsymbol{w}^T \boldsymbol{x}_A^i, i = 1, \ldots, N_A, \text{ and } y_B^j := \boldsymbol{w}^T \boldsymbol{x}_B^j, j = 1, \ldots, N_B$$

respectively. Means and covariances are computed from the training data. These empirical vector means are denoted $\boldsymbol{m}_A$ and $\boldsymbol{m}_B$ and the variance-covariance matrices are called S_A and S_B respectively. Let μ_A and μ_B be the (scalar) means for the values $\{y_A^i\}$ and $\{y_B^j\}$, and σ_A, σ_B the corresponding variances, and will depend on the choice of vector $\boldsymbol{w}$. What is the quantity that LDA seeks to maximise in order to achieve classification accuracy? Explain why that is a good choice. Discuss whether the quantity is invariant under scaling. *[5 marks]*

(c) A linear hyperplane is described by the equation $y = \boldsymbol{w} \cdot \boldsymbol{x} + b$. The decision boundary in the figure is the line (representing a hyperplane) for which $y = 0$ (labelled 0) and is perpendicular to $\boldsymbol{w}$. The two parallel hyperplanes that go through the support vectors (points with thickened edges) are indicated by the values $y = \pm 1$. Explain why a large margin is necessary for robust classification. From the geometry of the figure show that the size of the margin along the direction of $\boldsymbol{w}$ is $\frac{2}{}$. You may take $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ to be support vect

[7 marks]

(d) The max margin classifier for the training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ is obtained by solving an optimisation problem

$$\min_{\boldsymbol{w},b} \max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} \alpha_n \left(y_n(\boldsymbol{w}^T \boldsymbol{x}_n + b) - 1 \right), \ \alpha_n \geq 0.$$

Explain the motivation behind such an expression describing what each term stands for. *[6 marks]*

4

(a) For a probability distribution p over an event space $\mathcal{X}$ $p := \{p_i | i \in \mathcal{X}\}$, explain why the entropy $H(p)$

$$H(p) = \sum_i p_i \log(\frac{1}{p_i})$$

is often interpreted as the average degree of 'surprise'. *[3 marks]*

(b) For 2 probability distributions $p = \{p_i | i \in \mathcal{X}\}$ and $q = \{q_i | i \in \mathcal{X}\}$ over the same event space $\mathcal{X}$ the cross-entropy $H(p, q)$ and the Kullback-Leibler (KL) divergence $KL(p\|q)$ are defined as:

$$H(p, q) = \sum_i p_i \log(\frac{1}{q_i}) \text{ and } KL(p\|q) = H(p, q) - H(p) = \sum_i p_i \log(\frac{p_i}{q_i}).$$

Provide some intuition for what each of the two metrics over pairs of probability distributions captures. You may treat the distribution p as the one representing 'ground truth.' *[6 marks]*

(c) You are given a sequence $\boldsymbol{x} := \{x^{(1)}, \ldots, x^{(N)}\}$ of heads ($x^{(i)} = H$) and tails ($x^{(i)} = T$) which are the outcomes of N tosses of a (potentially biased) coin. All possible outcomes of N coin tosses would constitute the event space $\mathcal{X}$. A binomial distribution $B(N, \theta)$ sets the probability of occurrence of n_1 events of type 1, and $n_2 = N - n_1$ of type 2 as

$$P(n_1, n_2|N, \theta) = \frac{N!}{n_1! n_2!}\theta^{n_1}(1-\theta)^{n_2}.$$

Describe how you would fit the data $\mathcal{X}$ to a binomial distribution using maximum likelihood estimation and find the result to be the same as the empirical distribution $\tilde{p} = (\tilde{p}_H, \tilde{p}_T)$:

$$\theta^* = \frac{n_H}{N} =: \tilde{p}_H$$

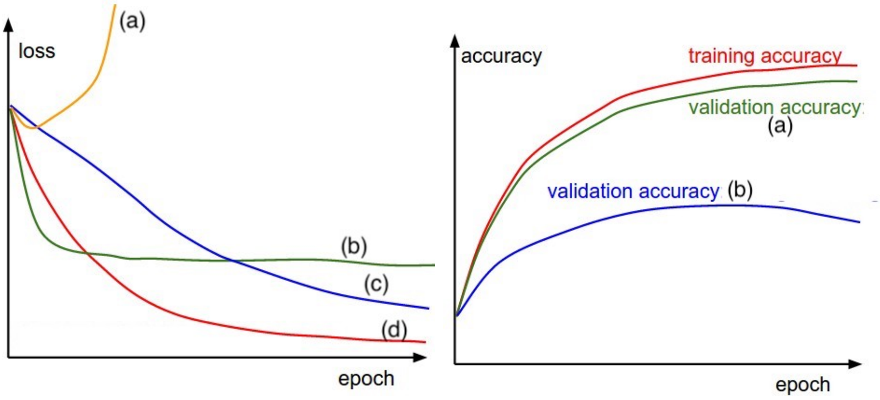
. *[10 marks]*

(d) A 2-dimensional Gaussian distribution is defined by the probability density function of $\boldsymbol{X} = (X_1, X_2)^T$ parameterised by its mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and variance-covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

$$p(\boldsymbol{X}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right).$$

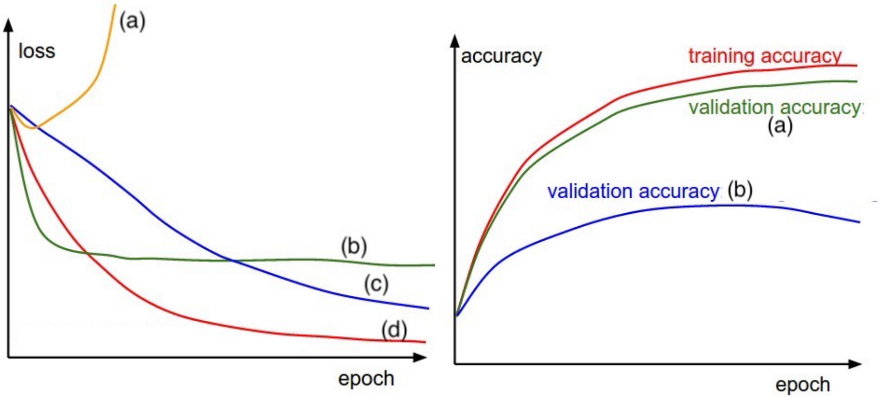
Draw 3 contour plots of equiprobable values of (X_1, X_2) for the cases (a) $\rho = 0$, (b) $\rho < 0$ and (c) $\rho > 0$ providing reasons for doing so. *[6 marks]*

5 Suppose we are training a neural network to do classification on the MNIST dataset. We run various experiments and plot the loss and accuracy over epochs and observe the plots shown in the figures below.



(a) In order to generate the loss vs. epochs plot above, the learning rate is modified four times resulting in the curves marked (a), (b), (c) and (d). The trends behave quite differently as a result of modifying the learning rate. You need to decide which learning rate to use. Which case would you choose (a), (b), (c) or (d) and explain why. *[6 marks]*

(b) Looking at the loss curve marked (a) in the loss plot, how can you modify your network to solve the issue? *[6 marks]*



(c) Now consider the accuracy of your model on training versus validation sets over epochs and assume you observe the results shown in the above figure. Comparing the curves labeled (a) and (b) which are both plotting the validation accuracy, which of these cases illustrates overfitting? Can you explain why? *[7 marks]*

(d) How can you prevent your neural network model from overfitting? *[6 marks]*

END OF PAPER