Problem 1 [15 marks]. Consider a logistic regression model $y = \sigma(\mathbf{w}^\top \mathbf{x} + b)$ and a two-way classification model $\mathbf{y} = \mathrm{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})$ for d-dimensional input $\mathbf{x} \in \mathbb{R}^d$.

a) [5 marks] Write out the formulas of the sigmoid and softmax functions.

b) [5 marks] How many model parameters do we have for the logistic regression model and the softmax regression model, respectively?

c) [5 marks] Given the same set of training data, which model (logistic vs softmax) is more likely to overfit? And why?

Hint: A d-dimensional vector counts d parameters. No derivation or proof is needed. The question is exactly the same as mid-term except the mark distribution.

a) $\sigma(z) = \frac{1}{1+e^{-z}}$

$[\mathrm{softmax}(z)]_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$

b): logistic: $d+1$

softmax: $2(d+1)$

c): same. because the hypothesis class, training objective, inference criterion are equivalent.

Problem 2 [20 marks]. In deep neural networks, a layer $\boldsymbol{h} \in \mathbb{R}^d$ may have multiple input layers $\boldsymbol{h}_1 \in \mathbb{R}^{d_1}, \boldsymbol{h}_2 \in \mathbb{R}^{d_2}$, calculated as $\boldsymbol{h} = f(W_1\boldsymbol{h}_1 + W_2\boldsymbol{h}_2 + \boldsymbol{b})$.

a) [10 marks] Show that this is equivalent to concatenating $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ as $\tilde{\boldsymbol{h}} = \binom{\boldsymbol{h}_1}{\boldsymbol{h}_2}$ and processing it by a neural layer $\boldsymbol{h} = f(\tilde{W}\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{b}})$.

b) [10 marks] Express $\tilde{W}$ and $\tilde{\boldsymbol{b}}$ in terms of $W_1, W_2, \boldsymbol{b}$. What are the dimensions of $\tilde{W}$ and $\tilde{\boldsymbol{b}}$?

a) $W_1 h_1 + W_2 h_2 = [W_1 \quad W_2] \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}$

Thus $h = f\left([W_1 \quad W_2] \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + b \right)$

b): $\tilde{W} = [W_1 \ W_2]$, $\tilde{b} = b$

$\tilde{W} \in \mathbb{R}^{d \times (d_1+d_2)}$ $\tilde{b} \in \mathbb{R}^d$

Problem 3 [15 marks]. In Coding Assignment 2, we have the "subtracting maximum" trick for implementing softmax regression.

a) [5 marks] What is the trick and why is it needed?

b) [10 marks] Prove that it is correct.

a) For Softmax $y_i = \dfrac{\exp(z_i)}{\sum_j \exp(z_j)}$, we actually implement

$$y_i = \frac{\exp(z_i - z_*)}{\sum_j \exp(z_j - z_*)} \quad \text{where } z_* = \max_i \{z_i\}$$

This makes the model more numerically stable. because exp may be very large for a positive number

b)

$$\frac{\exp(z_i - z_*)}{\sum_j \exp(z_j - z_*)} = \frac{\exp(z_i) / \exp(z_*)}{\sum_j \exp(z_j)/\exp(z_*)} = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

Problem 4 [25 marks]. Consider the max *a posteriori* inference for a K-way softmax regression: $\hat{t}(\mathbf{x}) = \text{argmax softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})$

a) [10 marks] Show that the set $S = \{\mathbf{x} : \hat{t}(\mathbf{x}) = k\}$ is convex for any $k = 1, \cdots, K$.

b) [5 marks] What is the benefit for *max a posteriori* inference?

c) [10 marks] Suppose c_{ij} is the cost for predicting the ith category for a sample of the jth category. Let π_k be the probability that the sample is of category k. What is the decision rule that minimizes the expected cost for this sample?

a)

Consider any $x_1, x_2 \in S$

① $w_k^T x_1 + b_k \geq w_{k'}^T x_1 + b_{k'}$ for $k' \neq k$

② $w_k^T x_2 + b_k \geq w_{k'}^T x_2 + b_{k'}$

λ·① + $(1-\lambda)$·②:

$w_k^T(\lambda x_1 + (1-\lambda)x_2) + b_k \geq w_{k'}^T(\lambda x_1 + (1-\lambda)x_2) + b_{k'}$

Thus, $\hat{t}(\lambda x_1 + (1-\lambda)x_2) = K$

implying that $\lambda x_1 + (1-\lambda)x_2 \in S$

b). maximizing the expected accuracy

c): If prediction is i

The expected cost is

$$\mathop{\mathbb{E}}_{j \sim \pi}[c_{ij}] = \sum_j \pi_j c_{ij}$$

Thus the decision rule is to minimize

$$\operatorname*{argmin}_i \sum_j \pi_j c_{ij}$$

Problem 5 [15 marks]. For a categorical variable $x \sim \mathrm{cat}(\pi_1, \cdots, \pi_K)$, we learn that the maximum likelihood estimation is simply counting (recall Gaussian Mixture Models). Now consider a set of samples $\{x^{(m)}\}_{m=1}^M$.

a) [5 marks] Write out the maximum likelihood estimation (i.e., the formula of counting).

b) [10 marks] The likelihood is $\mathcal{L}(\boldsymbol{\pi}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \pi_k^{1\{x^{(m)}=k\}}$, where $1\{\cdot\}$ is an indicator function. Prove that maximizing the log-likelihood yields the counting formula.

Hint: Note that we have a constraint $\pi_1 + \cdots + \pi_K = 1$. You may represent $\pi_1 = 1 - \pi_2 - \cdots - \pi_K$ and seek a closed-form solution. Alternatively, you may use the Lagrange multiplier method, if you know it; however, this is not expected or required.

a) $\pi_K = \dfrac{\sum_{m=1}^{M} \mathbb{1}\{x^{(m)}=k\}}{M} = \dfrac{M_k}{M}$ where $M_k = \sum_{m=1}^{M} \mathbb{1}\{x^{(m)}=k\}$

b). $\log \mathcal{L}(\pi) = \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{1}\{x^{(m)}=k\} \log \pi_k$

$= \sum_{k=1}^{K} M_k \log \pi_k$

$= \left(\sum_{k=2}^{K} M_k \log \pi_k\right) + M_1 \log(1-\pi_2-\cdots-\pi_K)$

For $k=2,\cdots K$:

$\dfrac{\partial \log}{\partial \pi_k} \mathcal{L}(\pi) = M_k \cdot \dfrac{1}{\pi_k} + M_1 \cdot \dfrac{1}{1-\pi_2-\cdots\pi_K} \cdot (-1) \overset{\text{set}}{=} 0$

Thus, $M_k(1-\pi_2-\cdots\pi_K) = M_1 \pi_k$

$\pi_k = \dfrac{M_k}{M_1}(1-\pi_2-\cdots-\pi_k) = \dfrac{M_k}{M_1}\pi_1$

Since $\pi_1+\cdots+\pi_K=1$, we have $\pi_1 + \sum_{k=2}^{K} \dfrac{M_k}{M_1}\pi_1 = 1$

$\pi_1 + \dfrac{M-M_1}{M_1}\pi_1 = 1$

implying that $\pi_1 = \dfrac{M_1}{M}$

For $k=2,\cdots,K$

$\pi_k = \dfrac{M_k}{M_1} \cdot \dfrac{M_1}{M} = \dfrac{M_k}{M}$

Thus, in general, the estimate is

$\hat{\pi}_k = \dfrac{M_k}{M}$

Problem 6 [10 marks]. As you may have realized, a student who has learned more machine learning knowledge will perform well for this exam, analogous to (a)___4___. A student who has a strong problem-solving ability will also perform well for this exam, analogous to (b)___5___.

Fill in the blanks with the following options: 1) underfitting, 2) overfitting, 3) having a larger hypothesis class, 4) having more training data, 5) better generalization, or 6) more regularization.

END OF THE EXAM

Scrap paper

- Additional scrap paper and answer sheets are available upon request.
- May be used as an answer sheet if you mark the problem ID clearly.
- Print your name and ID (number) on every answer sheet submitted (1 bonus mark).
- If answer sheets are detached, ask the instructor/TA to staple them.