20W-COMSCIM146 midterm

JACOB KAUFMAN

TOTAL POINTS

70 / 80

QUESTION 1

true/fasle 18 pts

1.1 1 **0 / 2**

✓ **- 2 pts** **InCorrect**

- 0 pts Correct

1.2 2 **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Incorrect

1.3 3 **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Incorrect

1.4 4 **2 / 2**

- 0 pts Correct

✓ **- 0 pts** **Incorrect but all are given points**

1.5 5 **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Incorrect

1.6 6 **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Incorrect

1.7 7 **0 / 2**

- 0 pts Correct

✓ **- 2 pts** **Incorrect**

1.8 8 **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Incorrect

1.9 9 **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Incorrect

QUESTION 2

multiple choice 28 pts

2.1 10 **4 / 4**

✓ **- 0 pts** **Correct**

- 1 pts (a) is not selected

- 1 pts (c) is not selected

- 1 pts (d) is not selected

- 1 pts (b) is selected

2.2 11 **4 / 4**

✓ **- 0 pts** **Correct**

- 1 pts a is selected

- 1 pts b is NOT selected

- 1 pts c is selected

- 1 pts d is selected

2.3 12 **3 / 4**

- 0 pts Correct

✓ **- 1 pts** **a) is not selected**

- 1 pts b) is not selected

- 1 pts c) is selected

- 1 pts d is not selected

2.4 13 **3 / 4**

- 0 pts Correct

- 1 pts a) is true

- 1 pts b) is true

- 1 pts c is false

✓ **- 1 pts** **d is true**

2.5 14 **4 / 4**

✓ **- 0 pts** **Correct**

- **1 pts** a) is false
- **1 pts** b is true
- **1 pts** c is false
- **1 pts** d is true

2.6 15 4 / 4

✓ - **0 pts** **Correct**
- **1 pts** a is not selected
- **1 pts** b is not selected
- **1 pts** c is selected
- **1 pts** d is selected

2.7 16 2 / 4

- **0 pts** Correct

✓ - **1 pts** **a is true**

✓ - **1 pts** **b is false**
- **1 pts** c is false
- **1 pts** d is false

QUESTION 3

decision tree 13 pts

3.1 a 2 / 2

✓ - **0 pts** **Correct**
- **0.5 pts** correct expression. should use the provided identities for final answer.
- **2 pts** wrong
- **1 pts** partially correct

3.2 b 4 / 5

- **0 pts** Correct
- **2 pts** incorrect H(Y|X1=1)
- **2 pts** incorrect H(Y|X1=0)

✓ - **1 pts** **correct (form of) H(Y|X1=0), H(Y|X1=1), wrong (form of) H(Y|X1)**
- **1 pts** wrong/missing IG(information gain)

3.3 c 5 / 5

✓ - **0 pts** **Correct**
- **1 pts** missing IG
- **2 pts** wrong H[y|x2=0]
- **2 pts** wrong H[y|x2=1]
- **1 pts** correct H[y|x2=1], correct H[y|x2=0], wrong H[y|x2]
- **1 pts** wrong form of H[y|x]

3.4 d 1 / 1

✓ - **0 pts** **Correct**
- **1 pts** wrong

QUESTION 4

MLE 8 pts

4.1 a 3 / 3

✓ - **0 pts** **Correct**
- **2 pts** Click here to replace this description.
- **1 pts** Click here to replace this description.
- **3 pts** Click here to replace this description.

4.2 b 3 / 3

✓ - **0 pts** **Correct**
- **1 pts** Click here to replace this description.
- **2 pts** Click here to replace this description.
- **3 pts** Click here to replace this description.

4.3 c 2 / 2

✓ - **0 pts** **Correct**
- **1 pts** Click here to replace this description.
- **2 pts** Click here to replace this description.

QUESTION 5

logistic regression 12 pts

5.1 a 2 / 2

✓ - **0 pts** **Correct**
- **2 pts** The answer is not correct.
- **1 pts** Your expression is correct, but your result of C is not correct.

5.2 b 5 / 5

✓ - **0 pts** **Correct**
- **5 pts** Wrong answer

5.3 c 1 / 2

- **0 pts** Correct
- **2 pts** wrong answer

✓ - **1 pts** **Your answer is partially correct (you know the answer should be the sum of the answer you get for (b)).**

- **0.5 pts** C varies for different n, so you cannot get N log C or C^N

5.4 d 3 / 3

✓ - **0 pts** **Correct**

- **3 pts** Wrong answer
- **2 pts** You did not show the equivalence.
- **1 pts** Your result of l(theta) is not correct.
- **1 pts** your answers are partially correct.

QUESTION 6

6 name 1 / 1

✓ - **0 pts** **Correct**

CM146: Introduction to Machine Learning **Winter 2020**

Midterm

Feb. 10^{th}, 2020

- Please do not open the exam unless you are instructed to do so.
- This is a closed book and closed notes exam.
- Everything you need in order to solve the problems is supplied in the body of this exam OR in a cheatsheet at the end of the exam.
- Mark your answers ON THE EXAM ITSELF. If you make a mess, clearly indicate your final answer (box it).
- For true/false questions, CIRCLE True OR False and provide a brief justification for full credit.
- Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one) and provide a brief justification if the question asks for one.
- If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.
- If you run out of room for your answer in the space provided, please use the blank pages at the end of the exam and indicate clearly that you've done so.
- Do NOT put answers on the back of any page of the exam.
- You may use scratch paper if needed (provided at the end of the exam).
- You have 1 hour 45 minutes.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Legibly write your name and UID in the space provided below to earn 2 points.

Name: Jacob Kaufman

UID: 204 929 264

Name and UID		/1
True/False		/18
Multiple choice		/28
Decision tree		/13
Maximum likelihood		/8
Logistic regression		/12
Total		/80

True/False (18 pts)

1. (2 pts) We are using linear regression to predict height from weight and age. The optimal value of the residual sum of squares (RSS) will change if we measure weight in pounds instead of kilograms for each instance.

(True) False

$J(\theta) = \sum_n (h_\theta(x_n) - y_n)^2$

Changing x_n will change $J(\theta)$

2. (2 pts) On a dataset that is not linearly separable, the 1-nearest neighbors classifier obtains zero training error.

(True) False

1-NN will always obtain zero training error, because a training data point will always have itself as the 1-NN.

3. (2 pts) As the number of training examples grows toward infinity, the probability that logistic regression will over-fit the training data goes to zero.

(True) False

As a probabilistic model, the more samples used to train the model, the more the model predicts the population of all possible examples.

4. (2 pts) The K-nearest neighbor algorithm often uses Euclidean distance as the default distance metric: $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$. Suppose we instead use a new distance metric: $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1$. The classification results will change as a result of this distance metric.

True (False)

this is equivalent to stepping parallel to the axes and evaluating the distance that way. This distance measure will scale equally between coordinates. When changing to Euclidean distance, so the 3 KNN classification does not change.

5. (2 pts) The training error of the perceptron never increases with each iteration of the perceptron algorithm.

True (False)

It can increase if the new hyperplane misclassifies a higher proportion of data. This is not guaranteed to not happen.

6. (2 pts) The value of x at which $f(x)$ attains its maximum is the same as the value of x at which $e^{f(x)}$ attains its maximum.

(True) False

$\frac{d}{dx} f(x) = f'(x)$

$\frac{d}{dx} e^{f(x)} = f'(x) e^{f(x)}$

Because $e^{f(x)} > 0$, the two expressions will achieve maxima at the same point.

7. (2 pts) You compute $x^* = argmin_x f(x)$ and find that $x^* = -\infty$. $f(x)$ is not convex.

(True) False

If $f''(x) > 0 \quad \forall x \in \mathbb{R}$ then $f(x)$ must have a finite minimum.

8. (2 pts) A decision tree learned with *MaxDepth* parameter set to ∞ always attains zero training error.

True (False)

It is possible that there can be no perfectly fitted decision tree.

9. (2 pts) Stochastic Gradient Descent is faster per iteration than Batch Gradient Descent.

(True) False

The algorithm is faster because it does not use as much data as batch GD.

Multiple choice (28 pts)

CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one)

10. (4 pts) What strategy can help reduce over-fitting in decision trees.

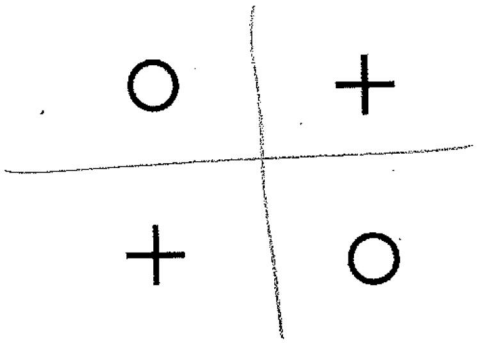
(a) Pruning

~~(b) Make sure it achieves zero training error~~

(c) Adding more training data

(d) Enforce a maximum depth for the tree

11. (4 pts) Consider the following data set: Circle the classifiers that will achieve zero training error on this data set.



~~(a) Logistic regression~~

(b) Depth-2 ID3 decision trees

~~(c) Perceptron~~

~~(d) 3NN classifier~~

12. (4 pts) Which of the following are true about linear regression?

~~(a) Its cost function is aways convex~~

(b) Its cost function is aways convex after adding l_2 regularization.

~~(c) The cost function always has a unique minimum~~.

(d) The cost function always has a unique minimum after adding l_2 regularization.

13. (4 pts) Consider a logistic regression model to predict if a yelp review is positive or not ($y = 1$ means the review is positive) based on two features: x_1 and x_2. x_1 is the number of times the word "great" appears and x_2 is the number of times the word "not" appears. The logistic regression model $P(y = 1|x; \boldsymbol{\theta}) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$ with $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2) = (2, 1, -2)$. Which of the following is true ?

(a) The decision boundary is given by the line $x_1 - 2x_2 + 2 = 0$

(b) If the word "great" appears more often (assuming everything else about the review is the same), probability that the review is classified as positive becomes closer to 1.

(c) If the word "not" appears more often (assuming everything else about the review is the same), probability that the review is classified as positive becomes closer to 1.

(d) If the review contains neither the word "great" nor the word "not", it will be classified as positive.

14. (4 pts) Which of the following is true of the Perceptron learning algorithm ?

(a) If the algorithm does not coverge within 100 iterations, the data is <u>not</u> linearly separable.

(b) If the algorithm coverges within 100 iterations, the data is linearly separable.

(c) If the data is linearly separable, it will coverge within 100 iterations.

(d) If the data is linearly separable, it may not converge in 100 iterations but may converge in 200 iterations.

15. The entropy of a distribution over a set of 3 items with probability mass function p is defined as $-\sum_{k=1}^{3} p(k) \log_2 p(k)$. Which of the following distributions has the lowest entropy?

(a) $(1, 0, 0)$

(b) $(0, 1, 0)$

(c) $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

(d) $(0.1, 0.3, 0.6)$

16. (4 pts) Given the same training data consisting of N instances $\{(x_n, y_n)\}_{n=1}^{N}$ where $x_n \in \mathbb{R}, y_n \in \mathbb{R}$, we fit two models: $h_1(x) = \theta_0 + \theta_1 x$ and $h_2(x) = \theta_0 + \theta_1 x + \theta_2 x^2$. For each model, we estimate its parameters by minimizing the residual sum of squares (RSS). Let RSS_m denote the minimum value of the residual sum of squares cost function evaluated on the training dataset for model m ($m \in \{1, 2\}$). Which of the following is always true ?

(a) $RSS_1 \geq RSS_2$

(b) $RSS_1 \leq RSS_2$

(c) $RSS_1 = RSS_2$

(d) RSS_1 can sometimes be greater and sometimes lesser than RSS_2.

Decision Tree (13 pts)

Suppose you want to build a decision tree for a problem. In the dataset, there are two classes (*i.e.*, Y can take one of two possible values), with 60 examples in the + class and 30 examples in the − class. Recall that the information gain for target label Y and feature X is defined as $Gain = H[Y] - H[Y|X]$, where $H[Y] = -E[\log_2 P(Y)]$ is the entropy. See cheatsheet at the end of this exam for entropy values.

(a) (2 pts) What is the entropy of the class variable Y?

$$H[Y] = -P(Y=0)\log P(Y=0) - P(Y=1)\log P(Y=1)$$
$$= -\frac{3}{9}\log\frac{3}{9} - \frac{6}{9}\log\frac{6}{9}$$
$$= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$
$$= B\left(\frac{1}{3}\right) = \boxed{0.92}$$

(b) (5 pts) For this data, we are interested in computing the information gain of a binary feature X_1. In the + class, the number of instances that have $X_1 = 0$ and $X_1 = 1$ respectively: $(30, 30)$. In the $-$ class, these numbers are: $(0, 30)$. Write down conditional entropy and information gain of X_1 relative to Y?

X_1	Y
+ x30	+ x60
− x30	
+ x30	− x30
− x0	

$P(Y=1|X_1=1) = \frac{30}{60} = \frac{1}{2}$

$P(Y=1|X_1=0) = 1$

$P(Y=0|X_1=1) = \frac{1}{2}$

$P(Y=0|X_1=0) = 0$

$\Rightarrow H[Y|X_1=1] = -P(Y=1|X_1=1)\log P(Y=1|X_1=1) - P(Y=0|X_1=1)\log P(Y=0|X_1=1)$

$= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = B(\frac{1}{2}) = 1$

$H[Y|X_1=0] = -P(Y=1|X_1=0)\log P(Y=1|X_1=0) - P(Y=1|X_1=0)\log P(Y=1|X_1=0)$

$= -1\log 1 - 0\log 0 = 0$

$\Rightarrow H[Y|X_1] = P(X_1=0)H[Y|X_1=0] + P(X_1=1)H[Y|X_1=1]$

$= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \boxed{\frac{1}{2}}$

$GAIN = H[Y] - H[Y|X] = 0.92 - 0.5 = \boxed{0.42}$

(c) (5 pts) We are interested in computing the information gain of a binary feature X_2. In the + class, the number of instances that have $X_2 = 0$ and $X_2 = 1$ respectively are: $(40, 20)$. In the $-$ class, these numbers are: $(20, 10)$. Write down conditional entropy and information gain of X_2 relative to Y?

X_2	Y
+ 20	+ ×60
− 40	
+ 10	− ×30
− 20	

$P(Y=1|X_2=1) = \frac{20}{30} = \frac{2}{3}$

$P(Y=1|X_2=0) = \frac{40}{60} = \frac{2}{3}$

$P(Y=0|X_2=1) = \frac{10}{30} = \frac{1}{3}$

$P(Y=0|X_2=0) = \frac{20}{60} = \frac{1}{3}$

$H[Y|X_2=1] = -P(Y=0|X_2=1)\log P(Y=0|X_2=1) - P(Y=1|X_2=1)\log P(Y=0|X_2=1)$

$= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = B(\frac{1}{3}) = 0.92$

$H[Y|X_2=0] = -P(Y=0|X_2=0)\log P(Y=1|X_2=0) - P(Y=1|X_2=0)\log P(Y=1|X_2=0)$

$= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = B(\frac{1}{3}) = 0.92$

From (a) $H[Y] = 0.92$

$H[Y|X_2] = P(X_2=0)\,H[Y|X_2=0] + P(X_2=1)\,H[Y|X_2=1]$

$= \frac{30}{60}(0.92) + \frac{30}{60}(0.92) = \boxed{0.92}$

$GAIN = H[Y] - H[Y|X_2] = 0.92 - 0.92 = \boxed{0}$

(d) (1 pts) Using information gain, which attribute will the ID3 decision tree learning algorithm choose as the root?

Because $GAIN_{X_1} > GAIN_{X_2}$,

ID3 will choose $\boxed{X_1}$ as the root.

Maximum Likelihood (8 pts)

Let $X_1, \ldots, X_N$ be i.i.d. random variables where $X_n \sim Exp(\lambda), n \in \{1, \ldots, N\}$. The probability density function for $X \sim Exp(\lambda)$ is:

$$p(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

(a) (3 pts) Give an expression for the log likelihood $l(\lambda)$ as a function of λ given this specific dataset. You may assume all values, $X_1, \ldots, X_N \geq 0$.

$$L(\lambda) = \prod_{n=1}^{N} p(x_n; \lambda) \Rightarrow \ell(\lambda) = \sum_{n=1}^{N} \log p(x_n; \lambda)$$

$$\Rightarrow \ell(\lambda) = \sum_{n=1}^{N} \log(\lambda e^{-\lambda x_n})$$

$$= \sum_{n=1}^{N} (\log \lambda + \log(e^{-\lambda x_n})) = \sum_{n=1}^{N} (\log \lambda + (-\lambda x_n) \log e)$$

$$= \sum_{n=1}^{N} (\log \lambda - \lambda x_n)$$

$$= N \log \lambda - \sum_{n=1}^{N} \lambda x_n \quad \boxed{= N \log \lambda - \lambda \sum_{n=1}^{N} x_n}$$

(b) (3 pts) Compute the derivative of the log likelihood for this specific dataset.

$$\ell(\lambda) = N \log \lambda - \lambda \sum_{n=1}^{N} x_n$$

$$\ell'(\lambda) = \boxed{\frac{N}{\lambda} - \sum_{n=1}^{N} x_n}$$

(c) (2 pts) What is the maximum likelihood estimate $\hat{\lambda}$ of λ?

$$\ell'(\lambda) = 0 \Rightarrow \frac{N}{\lambda} - \sum_{n=1}^{N} x_n = 0$$

$$\Rightarrow \sum_{n=1}^{N} x_n = \frac{N}{\lambda}$$

$$\Rightarrow \frac{\lambda}{N} = \frac{1}{\sum_{n=1}^{N} x_n}$$

$$\Rightarrow \lambda = \boxed{\hat{\lambda} = \frac{N}{\sum_{n=1}^{N} x_n}}$$

Alternative to logistic regression (12 pts)

In class, we discussed the logistic regression model for binary classification problem. Here, we consider an alternative model. We have a training set $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ where $\boldsymbol{x}_n \in \mathbb{R}^{D+1}$ and $y_n \in \{0, 1\}$. Like in logistic regression, we will construct a probabilistic model for the probability that y_n belongs to class 0 or 1, given $\boldsymbol{x}_n$ and the model parameters, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ ($\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \mathbb{R}^{D+1}$). More specifically, we model the target y_n as:

$$\begin{aligned} p(y_n = 0|\boldsymbol{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= Ce^{\boldsymbol{\theta}_0^\top \boldsymbol{x}_n} \\ p(y_n = 1|\boldsymbol{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= Ce^{\boldsymbol{\theta}_1^\top \boldsymbol{x}_n} \end{aligned} \tag{1}$$

(a) (2 pts) Find the value of C that makes Equation 1 a valid probability distribution for y_n.

$$p(y_n = 0 | x_n; \theta_0, \theta_1) = Ce^{\theta_0^T x_n}$$

$$p(y_n = 1 | x_n; \theta_0, \theta_1) = Ce^{\theta_1^T x_n}$$

$$\Sigma p = 1 \Rightarrow Ce^{\theta_0^T x_n} + Ce^{\theta_1^T x_n} = 1$$

$$\Rightarrow C\left(e^{\theta_0^T x_n} + e^{\theta_1^T x_n}\right) = 1$$

$$\Rightarrow \boxed{C = \frac{1}{e^{\theta_0^T x_n} + e^{\theta_1^T x_n}}}$$

(b) (5 pts) Write the log likelihood of the parameters for a single instance $(\boldsymbol{x}_n, y_n)$: $l(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \log p(y_n|\boldsymbol{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. Express your answer in terms of $y_n, \boldsymbol{x}_n, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. (Hint: use the notation in class where for a Bernoulli random variable Y_n that takes values 1 with probability θ and 0 with probability $1-\theta$, we can write its probability mass function: $p(y_n) = \theta^{y_n}(1-\theta)^{1-y_n}$).

$\ell(\theta_0, \theta_1) = \log p(y_n | x_n; \theta_0, \theta_1)$

We know from the hint

$p(y_n) = \theta^{y_n}(1-\theta)^{1-y_n}$

where θ is the probability $y_n = 1$.

$\theta = Ce^{\theta_1^T x_n}$. From (a),

$1-\theta = Ce^{\theta_0^T x_n}$

$\Rightarrow p(y_n) = (Ce^{\theta_1^T x_n})^{y_n}(Ce^{\theta_0^T x_n})^{1-y_n}$

$\Rightarrow \ell(\theta_0, \theta_1) = \log p(y_n | x_n; \theta_0, \theta_1)$

$= \boxed{\log\left((Ce^{\theta_1^T x_n})^{y_n}(Ce^{\theta_0^T x_n})^{1-y_n}\right)}$

where $C = \frac{1}{e^{\theta_0^T x_n} + e^{\theta_1^T x_n}}$

(c) (2 pts) Write the log likelihood of the parameters $\mathcal{LL}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ for the full training data.

From (c),

$$\log p(y_n | x_n; \theta_0, \theta_1) = \log\left((Ce^{\theta_1^T x_n})^{y_n} (Ce^{\theta_0^T x_n})^{1-y_n} \right)$$

$$= \log\left((Ce^{\theta_1^T x_n})^{y_n} \right) + \log\left((Ce^{\theta_0^T x_n})^{1-y_n} \right)$$

$$= \theta_1^T x_n y_n \log(C \cdot e) + \theta_0^T x_n (1-y_n) \log(C \cdot e)$$

$$= \theta_1^T x_n y_n (\log C + \log e) + \theta_0^T x_n (1-y_n)(\log C + \log e)$$

$$= \theta_1^T x_n y_n (1 + \log C) + \theta_0^T x_n (1-y_n)(1 + \log C)$$

$$= \left(\theta_1^T x_n y_n + \theta_0^T x_n (1-y_n) \right)(1 + \log C)$$

$$\mathcal{LL}(\theta_0, \theta_1) = \sum_{n=1}^{N} \log p(y_n | x_n; \theta_0, \theta_1)$$

$$= \sum_{n=1}^{N} (1 + \log C)\left(\theta_1^T x_n y_n + \theta_0^T x_n (1-y_n) \right)$$

$$= \boxed{(1 + \log C) \sum_{n=1}^{N} \left(\theta_1^T x_n y_n + \theta_0^T x_n (1-y_n) \right)}$$

where $C = \frac{1}{e^{\theta_1^T x_n} + e^{\theta_0^T x_n}}$

(d) (3 pts) Recall that in logistic regression, we model the target y_n as :

$$p(y_n = 0|\boldsymbol{x}_n; \boldsymbol{\theta}) = 1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n)$$
$$p(y_n = 1|\boldsymbol{x}_n; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_n)$$

Here $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Show that our new model (Equation 1) is exactly the same as the logistic regression model with $\boldsymbol{\theta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$.

these equations yield

$$p(y_n = 0 | x_n; \theta_0, \theta_1) = 1 - \sigma(\theta_1^T x_n - \theta_0^T x_n)$$

$$= 1 - \frac{1}{1+\exp(\theta_0^T x_n - \theta_1^T x_n)} = 1 - \frac{1}{1 + \frac{\exp(\theta_0^T x_n)}{\exp(\theta_1^T x_n)}}$$

$$= 1 - \frac{\exp(\theta_1^T x_n)}{\exp(\theta_1^T x_n) + \exp(\theta_0^T x_n)}$$

$$= \frac{\exp(\theta_0^T x_n)}{\exp(\theta_1^T x_n) + \exp(\theta_0^T x_n)}$$

$$= C e^{\theta_0^T x_n} \checkmark$$

and, $p(y_n = 1 | x_n; \theta_0, \theta_1) = \sigma(\theta_1^T x_n - \theta_0^T x_n)$

$$= \frac{1}{1 + \exp(\theta_0^T x_n - \theta_1^T x_n)} = \frac{1}{1 + \frac{\exp(\theta_0^T x_n)}{\exp(\theta_1^T x_n)}}$$

$$= \frac{\exp(\theta_1^T x_n)}{\exp(\theta_0^T x_n) + \exp(\theta_1^T x_n)} = C e^{\theta_1^T x_n} \checkmark$$

the probability distributions equal, so the models are the same. □

Identities

Probability density/mass functions for some distributions

$$\text{Normal} \quad : \quad P(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\text{Multinomial} \quad : \quad P(\boldsymbol{x}; \boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{x_k}$$

$\boldsymbol{x}$ is a length K vector with exactly one entry equal to 1 and all other entries equal to 0

$$\text{Poisson} \quad : \quad P(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

Matrix calculus

Here $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n \times n}$. $\boldsymbol{A}$ is symmetric.

$$\nabla \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}, \qquad \nabla \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{b}$$

Entropy

The entropy $H(X)$ of a Bernoulli random variable $X \sim \mathit{Bernoulli}(p)$ for different values of p:

p	$H(X)$
$\frac{1}{2}$	1
$\frac{1}{3}$	0.92
$\frac{1}{4}$	0.81
$\frac{1}{5}$	0.73
$\frac{2}{5}$	0.97

You may use this page for scratch space.

You may use this page for scratch space.