Homework 2

Logistic Regression, Decision Trees, Naïve Bayes

CMU 10-701: Introduction to Machine Learning (Fall 2018)

Authors: Chieh (Jessica) Lin, Jing Mao, Zirui Wang

OUT: Sept 19, 2018
DUE: **Oct 3, 2018, 2:59 PM**

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3.4"). Second, write your solution independently: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Submissions written in latex are preferred however handwritten submissions are also allowed, please follow the instructions below for each format:
  Latex: If you are submitting a Latex document each derivation/proof written between the begin{soln} and the end{soln} for that specific question.
  Handwritten: Submissions can be handwritten in which case please submit each solution on a seperate page. You are in charge of making sure your solutions are legible if we cannot read your solutions you will not be given credit for them.
  Upon submission, label each question using the template provided by Gradescope.
- **Programming**: All programming portions of the assignments should be submitted to Autolab. We will not be using this for autograding, meaning you may use any language which you like to submit.

1 Logistic Regression; Improving our understanding of Convexity (30 points)

Consider a binary classification problem where the goal is to predict a class $y \in \{0, 1\}$, given an input $x \in \mathcal{R}^p$. A method that you can use for this task is *Logistic Regression*. Recall that in *Logistic Regression*, the conditional log likelihood probability can be written as follows:

$$\mathcal{L}(w) = \log(y|\mathbf{X}, w) = \sum_{i=1}^{n} [y_i w^T x_i - \log\big(1 + \exp\big(w^T x_i\big)\big)]$$

where:

- $\mathbf{X} \in \mathcal{R}^{n \times (1+p)}$ is a data matrix, with the first column composed of all ones
- $w \in \mathcal{R}^{(p+1)\times 1}$ is the weight vector, with the first index w_1 acting as the bias term
- x_i is a column vector of the i^{th} row of $\mathbf{X}$
- $y \in \mathcal{R}^{n \times 1}$ is a column vector of labels $y_i \in \{0, 1\}$
- p is the dimension of data (number of features in each observation)

Our goal is to find the weight vector w that maximizes this likelihood. Unfortunately, for this model, we cannot derive a closed-form solution with MLE. An alternative way to solve for w is to use gradient ascent, and update w step by step towards the optimal w. But we know gradient ascent will converge to the optimal solution w that maximizes the conditional log likelihood $\mathcal{L}$ when $\mathcal{L}$ is concave. In this question, you will prove that $\mathcal{L}$ is indeed a concave function.

1. **[5 points]** A real-valued function $f : S \to \mathcal{R}$ defined on a convex set S, is said to be *convex* if,

   $$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2), \forall x_1, x_2 \in S, \forall t \in [0, 1].$$

   Show that a linear combination of n convex functions, $f_1, f_2, ..., f_n$, $\sum_{i=1}^{n} a_i f_i(x)$ is also a convex function $\forall a_i \in R^+$.

2. **[5 points]** Show that a linear combination of n concave functions, $f_1, f_2, ..., f_n$, $\sum_{i=1}^{n} a_i f_i(x)$ is also a concave function $\forall a_i \in R^+$. Recall that if a function $f(x)$ is convex, then $-f(x)$ is concave. (You can use the result from part (1))

3. **[5 points]** Another property of twice differentiable convex functions is that the second derivative is non-negative. Using this property, show that $f(x) = \log(1 + \exp x)$ is a convex function. Note that this property is both sufficient and necessary. i.e. (if $f''(x)$ exists, then $f''(x) \geq 0 \iff f$ is convex)

4. **[7 points]** Given two convex functions $f : \mathcal{R} \to \mathcal{R}$ and $g : \mathcal{R} \to \mathcal{R}$. Is their composition $g \circ f$ is also convex? If yes, prove it. If not, provide a counterexample and explain what additional conditions do we need to make $g \circ f$ convex. Please explain in detail. Note that $(g \circ f)(x) = g(f(x))$. (hint: the property given in part (3))

5. **[8 points]** Show that the log likelihood of *Logistic Regression* is a concave function.

2 Decision Trees; Improving our understanding of Information Theory (20 points)

In class we talked about entropy function and information gain, recall some definitions. For random variable X, Y that takes discrete values in $\{1, 2, ..., k\}$:
Entropy of X is

$$H(X) = -\sum_{i=1}^{k} p(X = i) \log_2 p(X = i)$$

Joint entropy of X,Y is

$$H(X, Y) = -\sum_{i=1}^{k} \sum_{j=1}^{k} p(X = i, Y = j) \log_2 p(X = i, Y = j)$$

Entropy of Y is conditioned on $X = x$

$$H(Y|X = x) = -\sum_{i=1}^{k} p(Y = i|X = x) \log_2 p(Y = i|X = x)$$

Conditional Entropy of Y given X

$$H(Y|X) = \sum_{i=1}^{k} p(X = i) H(Y|X = i)$$

Information gain between X and Y:

$$IG(X; Y) = H(X) - H(X|Y)$$

Information gain is also known as mutual information, and it is defined as follows:

$$\begin{aligned} I(X; Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

A way to measure the distance between two probability distribution is relative entropy, also known as Kullback-Leibler divergence

$$D(p(x)||q(x)) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

With the definitions above, prove the following. You can assume that the probability distribution is always positive for random variable X and Y. i.e. $p(X = i) > 0 \quad \forall i$.
(Hint: you may need to use Jensen's inequality for the questions below: $f(E[X]) \geq E[f(X)]$ when f is a concave function of X).

1. **[5 points]** Prove that $\log_2 k$ is an upper bound of entropy function for a discrete random variable with range $\{1, 2, ..., k\}$

2. **[7 points]** Suppose that $X \in \{1, 2, 3, 4\}$ is some discrete random variable. Suppose it has the following probability distribution: $P(X = 1) = 0.64$, $P(X = 2) = P(X = 3) = 0.16$ and $P(X = 4) = 0.04$. If we use a standard binary encoding for $X = 1$ to $X = 4$, we could represent $X = 1$ with code 00, $X = 2$ with code 01, $X = 3$ with code 10 and $X = 4$ with code 11. Then the expected bits per symbol is $(2 * 0.64 + 2 * 0.16 + 2 * 0.16 + 2 * 0.04) = 2$. This means that, on average, we need 2 bits to transmit X. However, we could reduce the number of bits we needed by using prefix-free encoding (where none of the codes are a prefix of the other codes). For example, $X = 1$ with code 0, $X = 2$ with code 10, $X = 3$ with code 110, $X = 4$ with code 1110; then the expected bits per symbol become $(1 * 0.64 + 2 * 0.16 + 3 * 0.16 + 4 * 0.04) = 1.6$. Show that entropy is a lower bound of expected bits per symbol given a prefix free encoding. Let $b(X = i)$ be the bits needed to represent the prefix-free encoding of $X = i$. You can assume the following property of a prefix-free encoding $\sum_{i=1}^{k} 2^{-b(X=i)} \leq 1$.

3. **[5 points]** Prove that relative entropy is non-negative

4. **[3 points]** From the result of the previous question, show that Information gain is non-negative, i.e. $IG(X; Y) \geq 0$

3 Naïve Bayes (20 points)

In this question, we use upper-case letters such as X, Y to denote random variables, and lower-case letters to denote values of random variables.

Suppose we let $X = (X_1, X_2, .., .X_n)$ denote the features, and $Y \in \{0, 1\}$ denote the label. Note that in any generative model approach, we model the conditional label distribution $P(Y|X)$ via the conditional distribution of features given the label $P(X|Y)$:

$$P(Y|X) \propto P(X|Y)P(Y). \tag{1}$$

1. **[1 point]** Rewrite the conditional distribution in (1) under the Naïve Bayes assumption that the features are conditionally independent given the label.

2. **[4 points]** Suppose that each feature X_i takes values in the set $\{1, 2, ..., K\}$. Further, suppose that the label distribution is Bernoulli, and the feature distribution conditioned on the label is multinomial. What is the total number of parameters of the model under the Naïve Bayes assumption? And without the Naïve Bayes assumption? Please give detailed step by step derivations.
   Suppose we change the set of values that Y takes, so that $Y \in \{0, 1, ..., M-1\}$. How would your answers change?

3. **[6 points]** Suppose each feature X_i takes values in the set $\{0, 1\}$. Suppose the label distribution is Bernoulli, and the feature distribution conditioned on label is also Bernoulli, with $\pi = P(Y = 1)$ and $\mu_{ijk} = P(X_i = k|Y = j)$, for $i = 1, 2, ..., n$, $k = 0, 1$ and $j = 0, 1$. Given N observations $\{(X^{(\ell)}, Y^{(\ell)})\}_{\ell=1}^{N}$, derive the MLE estimators of π and μ_{ijk} under the Naïve Bayes assumption.

4. Suppose each feature is real-valued, with $X_i \in \mathbb{R}$, and $P(X_i|Y = j) \sim \mathcal{N}(\theta_{ij}, 1)$ for $i = 1, 2, ..., n$ and $j = 0, 1$. Solve the following problems under the Naïve Bayes assumption.
   (a) **[3 points]** Given N observations $\{(X^{(\ell)}, Y^{(\ell)})\}_{\ell=1}^{N}$, derive the MLE estimator of θ_{ij}.

   (b) **[6 points]** Show that the decision boundary $\{(X_1, X_2, ..., X_n) : P(Y = 0|X_1, X_2, ..., X_n) = P(Y = 1|X_1, X_2, ..., X_n)\}$ is linear in $X_1, X_2, ..., X_n$.

4 Multiple Choice Questions (10 points)

- **There might be one or more right answers**. Please explain your choice in one or two sentences.

1. **[1 point]** Which of the following is **commonly** used to evaluate the performance of a logistic regression model?
   (A) Log loss
   (B) Accuracy
   (C) AUC-ROC
   (D) Mean Squared Error

2. **[2 points]** Which of the following are generative model based classifiers?

   (A) KNN
   (B) Naive Bayes
   (C) Logistic Regression
   (D) Linear Regression

3. **[2 points]** Suppose we learn a Naive Bayes model using MLE as an estimator, and where the inputs are binary with dimension two: $X \in \{0,1\}^2$, and the output is binary as well: $Y \in \{0,1\}$. Suppose the training data is $\{((0,0),0),((0,1),0),((1,1),0),((1,1),1),((1,0),1)\}$. What is the maximum likelihood estimation (MLE) of $P(Y=0|X=(1,0))$ under the Naive Bayes model? Show your steps.
   (A) $\frac{2}{5}$. (B) $\frac{1}{4}$. (C) $\frac{3}{5}$. (D) $\frac{11}{18}$. (E) Other (Specify your answer)

4. **[1 point]** Which of the following is the most suitable as an estimator of a logistic regression model?
   (A) Poission distribution
   (B) Ordinary least squares
   (C) Maximum likelihood estimation
   (D) Negative binomial distribution

5. **[2 points]** Indicate which of the following methods/models are used to solve a regression problem:
   (A) Logistic Regression
   (B) Linear Regression
   (C) Naive Bayes
   (D) KNN

6. **[2 points]** Which of the following statements are FALSE?
   (A) Decision tree is learned by minimizing information gain.
   (B) Maximizing the likelihood of logistic regression model yields multiple local optimums.
   (C) No classifier can do better than a Naive Bayes classifier if the distribution of the data is known.
   (D) The training error of 1-NN classifier is 0.

5 Programming Exercise (20 points)

Note: Your code for all of the programming exercises should also be submitted to Autolab. While visualizations and written answers should still be submitted to Gradescope as a part of the rest of the homework. In your code, **please use comments to points out primary functions that compute the answers to each question.**

Feel free to use any programming language, as long as your TAs can read your code. Turn in your code in a single .tar ball that might contain multiple source code files.

In this problem, you will implement the Naive Bayes (NB) algorithm on a pre-processed dataset that contains both **discrete** and **continuous** covariates. Recall from class that Naive Bayes classifiers assume the attributes $x^1, x^2, \ldots$ are conditionally independent of each other given the class label y, and that their prediction can be written as $\hat{y} = \text{argmax}_y P(y|X)$, where:

$$P(y|X = (x^1, \ldots, x^n)) \propto P(X, y) = P(X|y) \cdot P(y) = P(y) \cdot \prod_i P(x^i|y) \tag{2}$$

Consider the case where there are C classes, so that $y \in [C]$, and N different attributes.

- For a discrete attribute i that takes M_i different values, the distribution $P(x^i|y = c)$ can be modeled by parameters $\alpha_{i,c,1}, \alpha_{i,c,2}, \ldots, \alpha_{i,c,M_i}$, with $\sum_{j=1}^{M_i} \alpha_{i,c,j} = \sum_{j=1}^{M_i} P(x^i = j|y = c) = 1$. **Do NOT use smoothing**. Assume $\log(0) = \lim_{x \to 0} \log x = -\infty$.
- For a continuous attribute i, **in this question**, we can assume the conditional distribution is Gaussian; i.e. $P(x^i|y = c) = \mathcal{N}(\mu_{i,c}, \sigma^2_{i,c}) \approx \frac{1}{\sqrt{2\pi(\sigma^2_{i,c}+\varepsilon)}} \exp\left(-\frac{(x^i-\mu_{i,c})^2}{2(\sigma^2_{i,c}+\varepsilon)}\right)$, where $\mu_{i,c}$ and $\sigma^2_{i,c}$ are the mean and variance for attribute i given class c, respectively. In your implementation, you should estimate $\mu_{i,c}$ via the sample mean and $\sigma^2_{i,c}$ via the sample variance. **Meanwhile, take $\varepsilon = 10^{-9}$, which is a small value just to ensure the variance is not 0**.

You now need to implement a Naive Bayes algorithm that predicts whether a person makes over \$50K a year, based on various attributes about this person (e.g., age, education, sex, etc.). You can find the detailed description of the attributes, and download the data at

`http://mlr.cs.umass.edu/ml/datasets/Adult`.

You will need 2 files:

- `adult.data`[1]: Each line is a training data sample, with attributes listed in the same order as on the website and delimited by commas. For instance, the first entry of each line is **age**. The last entry of each line gives the correct label (>50K, ≤50K). There should be 32,561 training data samples.
- `adult.test`[2]: Same format as `adult.data`, but only used in evaluation of the model (i.e. testing), so you shouldn't use the label for training your NB classifier. There should be 16,281 testing data samples.

IMPORTANT: You should ignore (but do not delete) all incomplete data lines, which contains "?" as values for certain attributes in the line.

Hint: Because $P(y) \prod_i P(x^i|y)$ can get extremely small, you should use log-posterior for your computations:

$$\log\left[P(y)\prod_i P(x^i|y)\right] = \log P(y) + \sum_i \log P(x^i|y)$$

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
[2] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test

5.1 Report Parameters

For questions below, report only up to 4 significant digits after the decimal points.

(a) **[2 points]** Report the prior probability of each class.

(b) **[8 points]** For each class c, for each attribute i, print & report the following:

- If the attribute is discrete, report the value of $\alpha_{i,c,j}$ for every possible value j, **in the same order as on the website** (e.g., for attribute "sex", you should report the α for "Female" first, then "Male"). Clearly mark what the attribute is and what is the value of j.
- If the attribute is continuous, report the value of $\mu_{i,c}$ and $\sigma_{i,c}$.

For instance, your answer could have the following format (the values are made up):

(1) Class "$> 50K$":

- age: mean=43.9591, var=105.4513
- workclass: Private=0.02, Self-emp-not-inc=0.0134, Self-emp-inc=0.0998, ...
- ...
- native-country: ...

(2) Class ...

(c) **[2 points]** Report the log-posterior values (i.e. $\log[P(X|y)P(y)]$) for the first 10 test data (in the same order as the data), each rounding to 4 decimal places (have 4 numbers after decimal points, for example, 12.3456). Ignore the lines which contain "?" and report the values with the corresponding line numbers.

5.2 Evaluation

(a) **[1 point]** Evaluate the trained model on the training data. What is the accuracy of your NB model? Round your answer to 4 decimal places.

(b) **[1 point]** Evaluate the trained model on the testing data. What is the accuracy of your NB model? Round your answer to 4 decimal places.

(c) **[6 points]** Instead of training the NB using all training data, try to train only with the first n data [3], and then evaluate on the testing dataset. Report the testing accuracies for $n = \{2^i \text{ for } i = 5, \ldots, 13\}$ (i.e. $n = 32, \ldots, 8192$). Plot the training and testing accuracy vs. # of training data. What do you observe? At what value of n do testing accuracy and training accuracy attain maximum? **In general**, what would you expect to happen if we use only a few (say $n < 3$) training data for Naive Bayes? Explain briefly (hint: we did not use smoothing).

[3] The count includes those lines with "?", but you should ignore those lines when training.