CS 4780/5780 Homework 4

Due: Thus 10/18/18 11:55pm on Gradescope

Problem 1: Intuition for Naive Bayes

Ronnie is playing a game named Flippin' Extravaganza, he asks if someone can help him win. Can you help him beat the house?

a)

This game is easy to undertake, there is a red hat and a blue hat with a weighted penny respectively. The operator secretly flips one of them and asks you to guess under which hat it is. Suppose you know that the red hat's penny is weighted to come up heads 3/5 of the time, and the blue hat's penny is weighted to come up heads 7/10 of the time. If the penny comes up heads, what is the probability that it came from the red hat?

b)

To get you hooked, the operator next makes the game more complex, but with better odds. Each hat actually has a penny, a nickel, a dime, and a quarter – all weighted. The operator secretly selects all the coins from one hat at random, flips them all, and asks you to guess which hat they came from. Suppose you know the the red hat's penny, nickel, dime, and quarter come up heads with probability $[3/5, 3/10, 1/2, 4/5]$, respectively, and the blue hat's coins have heads probabilities $[7/10, 1/5, 1/10, 2/5]$, respectively. If the coins come up $[H, H, T, H]$, what is the probability that they came from the red hat?

Problem 2: Linearity of Gaussian Naive Bayes

In this question, you will show that Naive Bayes is a linear classifier when using Gaussian likelihood factors with shared variances. Specifically, consider the following Naive Bayes model:

$$p\left(y|\mathbf{x}\right) = \frac{\prod_{\alpha=1}^{d} p\left([\mathbf{x}]_\alpha|y\right) p\left(y\right)}{p\left(x\right)}$$

with:

$$p\left([\mathbf{x}]_\alpha|y\right) = \mathcal{N}\left([\mu_y]_\alpha, [\sigma]_\alpha\right)$$

That is, there is a separate mean value for each feature $\mathbf{x}_\alpha$ and each class $y \in \{0, 1\}$. However, variances are shared across classes, so that there is only one variance $[\sigma]_\alpha$ per feature.

a)

Show that the decision rule $p(y = 1|x)$ can equivalently be written as:

$$p(y = 1|\mathbf{x}) = \frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 1)p(y = 1)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 1)p(y = 1) + \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 0)p(y = 0)}$$

Hint: remember the sum rule and the product rule.

b)

Using this formulation, show how to rewrite $p(y = 1|x)$ as:

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-\log\frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=1)p(y=1)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=0)p(y=0)}\right)}$$

c)

Given the above expression for $p(y = 1|x)$, show that Naive Bayes with this definition of $p([\mathbf{x}]_\alpha|y)$ is a linear model. Hint: the form you derived in part b should remind you of a decision rule you have seen before.

Problem 3: Gradient for Logistic Regression

In this problem, we are going to assume the same notation setup in class. For logistic regression, we model the class probability by

$$P(y|\mathbf{x}_i) = \sigma(y(\mathbf{w}^T\mathbf{x}_i))$$

where we define

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

(Note: We dropped the bias term b since we can always absorb the bias into $\mathbf{w}$)

1. Show that the sigmoid function $\sigma(\cdot)$ has the following property

   $$\sigma(-s) = 1 - \sigma(s)$$

   By proving this property, we have shown that we have properly defined a probabilistic model, namely, $P(y_i = 1|\mathbf{x_i}) + P(y_i = -1|\mathbf{x_i}) = 1$

2. In class, we mentioned about using Gradient Descent to find the MLE estimate for $\mathbf{w}$. Here we are going to compute the gradient of the log likelihood function.

   (a) To make things easier, first show that

   $$\sigma'(s) = \sigma(s)(1 - \sigma(s))$$

   (b) Show that the gradient of the log likelihood function, namely, $\nabla_w \log P(\mathbf{y}|X, \mathbf{w})$ where $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ and $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ is

   $$\sum_{i=1}^{n}(1 - \sigma(y_i(\mathbf{w}^T\mathbf{x}_i)))y_i\mathbf{x}_i$$

Problem 4: Optimization with Gradient Descent

1. You have a univariate function you wish to minimize, $f(w) = 5(w - 11)^4$. Suppose you wish to perform gradient descent with constant step size $\alpha = 1/40$. Starting with $w_0 = 13$, perform 5 steps of gradient descent. What are $w_0, ..., w_5$? What is the value of $f(w_5)$?

2. With the same function and starting point above, perform gradient descent until convergence, with the constant step size $\alpha = 1/80$. What are all of the $w_0, w_1, ...$ and all of the $f(w_0), f(w_1), ...$ until w achieves the minimum point?

Problem 5: Linear Regression

Consider we have the following 1-d training set:

x	y
-2	7
-1	4
0	3
1	4
2	7

and our goal is to find a regression model that could regress x to our target value y. To do this, we are going to use a linear regression model. Namely, we are going to model our data by assuming the relationship

$$\begin{aligned} y &= w_1 x + w_0 + \epsilon \\ &= \mathbf{w}^T \phi(x) + \epsilon \end{aligned}$$

where $\phi(x) = [1, x]^T$. We call ϕ a feature mapping of x and this feature mapping allows us to absorb the bias w_0 into the vector $\mathbf{w}$.

1. With this feature mapping, we can write down the data matrix as

   $$X = [\phi(x_1)...\phi(x_n)]$$

   Using the formula given in class, compute the closed form solution for $\mathbf{w}$. (If you did everything correctly, the matrix you need to invert should be diagonal.)

2. Recall that the loss function for linear regression is

   $$\ell(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \phi(x_i))^2$$

   With the closed formed solution obtained in the previous question, calculate the training loss.