Runqiu Ye
Stanford CS299
Problem Set #1
06/23/2024

Problem Set #1: Supervised Learning

Problem 1 Linear Classifiers (Logistic Regression and GDA)
Consider two datasets provided in the following files:

i. `data/ds1_{train,valid},csv`

ii. `data/ds2_{train,valid},csv`

Each file contains m examples, one example per row. The i-th row contains columns $x_0^{(i)} \in \mathbb{R}$, $x_1^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \{0, 1\}$. Use logistic regression and GDA to perform binary classification.

(a) Average empirical loss for logistic regression:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log\big(h_\theta(x^{(i)})\big) + (1 - y^{(i)}) \log\big(1 - h_\theta(x^{(i)})\big),$$

where $y^{(i)} \in \{0, 1\}$, $h_\theta(x^{(i)}) = g(\theta^T x)$ and $g(z) = 1/(1 + e^{-z})$.

The gradient of the function

$$\frac{\partial J}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}.$$

It follows that

$$\frac{\partial^2 J}{\partial \theta_k \partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))x_k^{(i)}x_j^{(i)}.$$

Hence, The Hessian H of this function is

$$H = \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))x^{(i)}(x^{(i)})^T.$$

Now, for any vector z, using Einstein's summation, we have

$$\begin{aligned} z^T H z &= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))z_k x_k^{(i)} x_j^{(i)} z_j \\ &= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))(x^T z)^2 \\ &\geq 0 \end{aligned}$$

This shows that H is PSD, and J is convex.

(b) **Coding problem.**

(c) To show that GDA results in a classifier that has a linear decision boundary, we want to show

$$p(y = 1 \mid x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-(\theta^T x + \theta_0))}$$

for some $\theta \in \mathbb{R}^n$ and $\theta_0 \in \mathbb{R}$ as functions of ϕ, Σ, μ_0, and μ_1. We have

$$\begin{aligned}
p(y = 1 \mid x) &= \frac{p(x \mid y = 1)p(y = 1)}{p(x \mid y = 1)p(y = 1) + p(x \mid y = 0)p(y = 0)} \\
&= \frac{\phi \exp\big(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\big)}{\phi \exp\big(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\big) + (1 - \phi) \exp\big(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\big)} \\
&= \frac{1}{1 + \frac{1-\phi}{\phi} \exp\big(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\big)} \\
&= \frac{1}{1 + \frac{1-\phi}{\phi} \exp\big(-((\mu_1 - \mu_0)^T \Sigma^{-1} x + \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1))\big)}.
\end{aligned}$$

This is the desired form, where

$$\begin{aligned}
\theta &= \Sigma^{-1}(\mu_1 - \mu_0), \\
\theta_0 &= \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) - \log \frac{1 - \phi}{\phi}.
\end{aligned}$$

(d) The log-likelihood of the data is

$$\begin{aligned}
\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \\
&= \sum_{i=1}^{m} 1\{y^{(i)} = 1\} \left(-\frac{1}{2}(x^{(i)} - \mu_1)^T \Sigma^{-1}(x^{(i)} - \mu_1) + \log \phi \right) \\
&\quad + \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left(-\frac{1}{2}(x^{(i)} - \mu_0)^T \Sigma^{-1}(x^{(i)} - \mu_0) + \log(1 - \phi) \right) \\
&\quad - \frac{m}{2} \log |\Sigma| + C,
\end{aligned}$$

where C is some constant independent of the parameters.

Let $\nabla_\phi \ell = 0$, we have

$$\phi = \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)} = 1\}.$$

Let $\nabla_{\mu_1} \ell = 0$, we have

$$\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \Sigma^{-1} x^{(i)} = \sum_{i=1}^{m} 1\{y^{(i)} = 1\} \Sigma^{-1} \mu_1,$$

and thus

$$\mu_1 = \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}}, \quad \mu_0 = \frac{\sum_{i=1}^m 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}}.$$

To derive Σ, recall that $\nabla_A \log |A| = (A^{-1})^T$, so we have

$$\nabla_{\Sigma^{-1}} \ell = -\frac{m}{2}\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

Hence,

$$\Sigma = \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

We conclude that the maximum likelihood estimates of the parameters are given by

$$\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^m 1\{y^{(i)} = 1\}, \\
\mu_0 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}}, \\
\mu_1 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}}, \\
\Sigma &= \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{iy})^T.
\end{aligned}$$

(e) **Coding problem.**

(f) See jupyter notebook for plots.

(g) See jupyter notebook for plots. On Dataset 1 GDA perform worse than logistic regression. This might be the case because for Dataset 1, the distribution of features are not quite multivariate normal.

(h) *** TO-DO ***

■

Problem 2 Incomplete, Positive-Only Labels

Dataset without full access to labels. In particular, we have labels only for a subset of positive examples. All negative examples and the rest of positive examples are unlabeled. Assume dataset $\{(x^{(i)}, t^{(i)}, y^{(i)})\}_{i=1}^m$ where $t^{(i)} \in \{0, 1\}$ is true label and where

$$y^{(i)} = \begin{cases} 1 & x^{(i)} \text{ is labeled} \\ 0 & \text{otherwise.} \end{cases}$$

All labeled examples are positive, which is to say $p(t^{(i)} = 1 \mid y^{(i)} = 1) = 1$. Goal is to construct a binary classifier h of true label t which only access to partial labels y. That is, construct h such that $h(x^{(i)}) \approx p(t^{(i)} = 1 \mid x^{(i)})$ as closely as possible, using only x and y.

(a) Suppose each $y^{(i)}$ and $x^{(i)}$ conditionally independent given $t^{(i)}$:

$$p(y^{(i)} = 1 \mid t^{(i)} = 1, x^{(i)}) = p(y^{(i)} = 1 \mid t^{(i)} = 1).$$

That is, labeled examples are selected uniformly at random from positive examples.

Want to show $p(t^{(i)} = 1 \mid x^{(i)}) = p(y^{(i)} = 1 \mid x^{(i)})/\alpha$ for some $\alpha \in \mathbb{R}$. As $p(\cdot \mid x^{(i)})$ is a conditional measure, we have

$$\begin{aligned} p(y^{(i)} = 1 \mid x^{(i)}) &= p(y^{(i)} = 1 \mid t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 \mid x^{(i)}) \\ &\quad + p(y^{(i)} = 1 \mid t^{(i)} = 0, x^{(i)}) p(t^{(i)} = 0 \mid x^{(i)}) \\ &= p(y^{(i)} = 1 \mid t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 \mid x^{(i)}) \\ &= p(y^{(i)} = 1 \mid t^{(i)} = 1) p(t^{(i)} = 1 \mid x^{(i)}). \end{aligned}$$

Hence, $p(t^{(i)} = 1 \mid x^{(i)}) = p(y^{(i)} = 1 \mid x^{(i)})/\alpha$ where $\alpha = p(y^{(i)} = 1 \mid t^{(i)} = 1)$.

(b) Estimate α using a trained classifier h and a held-out validation set V. Let $V_+ = \{x^{(i)} \in V \mid y^{(i)} = 1\}$. Assuming $h(x^{(i)}) \approx p(y^{(i)} = 1 \mid x^{(i)})$ for all $x^{(i)}$. Want to show

$$h(x^{(i)}) \approx \alpha \text{ for all } x^{(i)} \in V_+.$$

May assume that $p(t^{(i)} = 1 \mid x^{(i)}) \approx 1$ when $x^{(i)} \in V_+$.

We have

$$\begin{aligned} h(x^{(i)}) &\approx p(y^{(i)} = 1 \mid x^{(i)}) \\ &= p(y^{(i)} = 1 \mid t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 \mid x^{(i)}) \\ &\approx \alpha. \end{aligned}$$

(c) **Coding problem.**

(d)

■