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:

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

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

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

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) **Coding problem.**

(e) **Coding problem.** Estimate the constant α using validation set.

$$\alpha \approx \frac{1}{|V_+|} \sum_{x^{(i)} \in V_+} h(x^{(i)}).$$

To plot the decision boundary, we need to calculate the rescaled θ, write θ_*. The new decision boundary is given by $\frac{1}{\alpha} \frac{1}{1+\exp(-\theta^T x)} = \frac{1}{2}$. We have

$$\theta^T x + \log\left(\frac{2}{\alpha} - 1\right) = 0.$$

This is equivalent to $\theta_*^T x = 0$. This shows that θ_* and θ differs only in the 0-th index by a constant $\log\left(\frac{2}{\alpha} - 1\right)$.

■

Problem 3 Poisson Regression

(a) The poisson distribution parametrized by λ is

$$p(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}.$$

Therefore, we have

$$p(y; \lambda) = \frac{1}{y!} \exp(-\lambda + y \log \lambda).$$

Compare with $p(y; \eta) = b(y) \exp\big(\eta^T T(y) - a(\eta)\big)$, we conclude that the poisson distribution is in the exponential family, with

$$\begin{aligned}
b(y) &= \frac{1}{y!}, \\
T(y) &= y, \\
\eta &= \log \lambda, \\
a(\eta) &= e^\eta.
\end{aligned}$$

(b) The canonical response function for the family

$$\mathbb{E}[T(y); \eta] = \mathbb{E}[T(y); \eta] = \lambda = e^\eta.$$

(c) For a general linear model and a training set, the log likelihood

$$\begin{aligned}
\log p(y^{(i)} \mid x^{(i)}; \eta) &= \log b(y) \exp\big(\eta^T T(y) - a(\eta)\big) \\
&= \log b(y) + \eta^T T(y) - a(\eta).
\end{aligned}$$

For our model with poisson responses y, we have

$$\ell = \log p(y^{(i)} \mid x^{(i)}; \theta) = -\log y! + (\theta^T x^{(i)}) y^{(i)} - \exp\big(\theta^T x^{(i)}\big).$$

Taking the derivative with respect to θ_j, we have

$$\frac{\partial \ell}{\partial \theta_j} = \big(y^{(i)} - \exp\big(\theta^T x^{(i)}\big)\big) x_j^{(i)}$$

Hence, the stochastic gradient ascent update rule for learning using a GLM model with poisson response y is

$$\begin{aligned} \theta_j &:= \theta_j + \alpha \frac{\partial \ell}{\partial \theta_j} \\ &:= \theta_j + \alpha \big(y^{(i)} - \exp\big(\theta^T x^{(i)}\big)\big) x_j^{(i)}. \end{aligned}$$

(d) **Coding problem.** To predict the dataset, recall that the hypothese function for our model with poisson response y is

$$h_\theta(x) = \mathbb{E}[y \mid x] = e^\eta = e^{\theta^T x}.$$

Also, for the model, we utilize batch gradient ascent:

$$\theta_j := \theta_j + \frac{\alpha}{m} \sum_{i=1}^{m} \big(y^{(i)} - \exp\big(\theta^T x^{(i)}\big)\big) x_j^{(i)}.$$

■

Problem 4 Convexity of Generalized Linear Models
Investigate nice properties of GLM. Goal is to show that the negative log-likelihood (NLL) loss of a GLM is convex with respect to the model parameters.
Recall that for exponential family distribution

$$p(y; \eta) = b(y) \exp\big(\eta^T T(y) - a(\eta)\big),$$

where η is the *natural parameter* of distribution. Our approach is to show the Hessian of loss w.r.t the model parameters is PSD.
Restrict to the case where η is scalar and η is modeled as $\theta^T x$. Assume $p(Y \mid X; \theta) \sim$ ExponentialFamily(η) where $\eta \in \mathbb{R}$ is a scalar and $T(y) = y$. That is

$$p(y; \eta) = b(y) \exp(\eta y - a(\eta)).$$

(a) The mean of the distribution

$$\mathbb{E}[y; \eta] = \int y p(y; \eta) dy = \int y b(y) \exp(\eta y - a(\eta)) dy.$$

Following the hint, observe that

$$\begin{aligned}\frac{\partial}{\partial\eta}\int p(y;\eta)dy &= \int \frac{\partial}{\partial\eta}p(y;\eta)dy\\ &= \int b(y)\left(y-\frac{\partial a}{\partial\eta}\right)\exp(\eta y-a(\eta))dy.\end{aligned}$$

While $\int p(y;\eta)dy = 1$, we have $\frac{\partial}{\partial\eta}\int p(y;\eta) = 0$ and

$$\mathbb{E}[y;\eta] = \int b(y)\frac{\partial a(\eta)}{\partial\eta}\exp(\eta y-a(\eta))dy.$$

Since $\frac{\partial a(\eta)}{\partial\eta}$ does not depend on y, we have

$$\mathbb{E}[y;\eta] = \int b(y)\frac{\partial a(\eta)}{\partial\eta}\exp(\eta y-a(\eta))dy = \frac{\partial a(\eta)}{\partial\eta}\int b(y)\exp(\eta y-a(\eta))dy = \frac{\partial a(\eta)}{\partial\eta}.$$

This shows that $\mathbb{E}[Y \mid X;\theta]$ can be represented as the gradient of the log-partition function a with respect to the natural parameter η.

(b) Notice that

$$\begin{aligned}\frac{\partial\mathbb{E}[y;\eta]}{\partial\eta} &= \frac{\partial}{\partial\eta}\int yb(y)\exp(\eta y-a(\eta))dy\\ &= \int yb(y)\left(y-\frac{\partial a}{\partial\eta}\right)\exp(\eta y-a(\eta))dy\\ &= \int y^2b(y)\exp(\eta y-a(\eta))dy - \int yb(y)\frac{\partial a}{\partial\eta}\exp(\eta y-a(\eta))dy\\ &= \mathbb{E}[y^2;\eta]-\frac{\partial a}{\partial\eta}\mathbb{E}[y;\eta]\\ &= \mathbb{E}[y^2;\eta]-(\mathbb{E}[y;\eta])^2\\ &= \mathrm{Var}(y;\eta).\end{aligned}$$

This completes the proof, and we can see that $\mathrm{Var}(Y \mid X;\theta)$ can be expressed as the second derivative of the mean w.r.t η (i.e. the second derivative of log-partition function $a(\eta)$ w.r.t natural parameter η).

(c) The loss function $\ell(\theta)$, the NLL of the distribution

$$\begin{aligned}\ell(\theta) &= -\log\prod_{i=1}^m p(y^{(i)} \mid x^{(i)};\eta)\\ &= -\sum_{i=1}^m \log p(y^{(i)} \mid x^{(i)};\eta)\\ &= \sum_{i=1}^m -\log b(y^{(i)}) - \eta y^{(i)} + a(\eta)\\ &= \sum_{i=1}^m -\log b(y^{(i)}) - y^{(i)}\theta^T x^{(i)} + a(\theta^T x^{(i)}).\end{aligned}$$

Now, to calculate the Hessian of the loss function w.r.t θ, we first calculate

$$\frac{\partial \ell}{\partial \theta_k} = \sum_{i=1}^{m} \left(\frac{\partial a}{\partial \eta} - y^{(i)} \right) x_k^{(i)}.$$

It follows that

$$\frac{\partial \ell}{\partial \theta_j \theta_k} = \sum_{i=1}^{m} \frac{\partial^2 a}{\partial \eta^2} x_j^{(i)} x_k^{(i)}.$$

Hence, the Hessian of the loss function is

$$H = \sum_{i=1}^{m} \frac{\partial^2 a}{\partial \eta^2} x^{(i)} (x^{(i)})^T.$$

To prove the Hessian is always PSD, consider any $z \in \mathbb{R}^n$, where n is the dimension of $x^{(i)}$, and

$$\begin{aligned} z^T H z &= \sum_{i=1}^{m} z_j H_{jk} z_k \\ &= \sum_{i=1}^{m} \frac{\partial^2 a}{\partial \eta^2} z_j x_j x_k z_k \\ &= \sum_{i=1}^{m} \mathrm{Var}(Y \mid X; \eta) (x^T z)^2 \\ &\geq 0, \end{aligned}$$

since the variance is always non-negative. This completes the proof that NLL loss of GLM is convex.

■

- Any GLM model is *convex* in its model parameters.
- The exponential family of probability distribution are mathematically nice. We can caucluate the means and variance using derivatives, which is easier that integrals.

△

Problem 5 Locally weighted linear regression

(a) Weighted linear regression. Specifically, want to minimize

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$$

i. Let X be the m by n matrix where the i-th row is $(x^{(i)})^T$, and let y be the m by 1 matrix where the i-the row is $y^{(i)}$. Then J can also be written

$$J(\theta) = (X\theta - y)^T W(X\theta - y),$$

where W is the diagonal matrix

$$W_{ij} = \frac{1}{2}\delta_{ij} w^{(i)}.$$

ii. If all the $w^{(i)}$ is 1, then the normal equation is

$$X^T X\theta = X^T y,$$

and the value of θ that minimizes $J(\theta)$ is given by $(X^T X)^{-1} X^T y$. Here, to genrealize the normal equation, we first calculate the derivative

$$\nabla_\theta J = X^T(2W(X\theta - y)) = 2X^T W X\theta - 2X^T W y.$$

Setting this to 0, we get the normal equation

$$X^T W X\theta = X^T W y$$

and the expression for θ

$$\theta = (X^T W X)^{-1} X^T W y.$$

Notice that for $w^{(i)} = 1$, $W = I$ and we get the original form of the normal equation.

iii.

■