CS229 Problem Set #2

Supervised learning II

1) Logistic Regression: Training Stability

a) The model didn't converge on the dataset B.

b) The model provided tries to find a linear decision boundary with mapping the features to a higher dimension.

It does so by calculating the functional margins of the data $\hat{\gamma}^{(i)} = y^{(i)}(\theta^T x^{(i)} + b)$.

For each correct classification $\hat{\gamma}^{(i)} > 0$

Also, in the implementation provided the margin is inversly proportional to the gradient.

```
probas = Y * X.dot(theta)
grad = -(1/m) * (X.T.dot(probas * Y))
```

Since the dataset B is not linearly seperable, θ keeps changing considerably (the convergence condition over θ is not attained) after each epoch.

c) i) The convergence condition in the Provided model is ~~that~~ prev_theta - theta $< 1e-15$

Seting a larning rate $\alpha << 1e^{-15}$ could cause the model to converge even after the first update. But the θ is not optimal and thus not the goal of the model.

ii) Decreasing the larning rate by $1/t^2$ seems to be a reasonable solution, which makes the algorithm to converge.
But it is not guaranteed that we have an ophimal estimation θ.

iii) We can nohice that the dataset is already linearly scaled.

iv) Adding L2 regularisation would would reduce $\|\theta\|^2$ and thus ~~reduce~~ the increase the margins in miss calassification

2. Model Calibration

A binary classification model is said to be well calibrated given a range (a,b) if

* $0 \leq a < b \leq 1$

* $\sum_{i \in I_{a,b}} P(y^{(i)} = 1 | x^{(i)}; \theta) = \sum_{i \in I_{a,b}} \mathbb{I}\{y^{(i)} = 1\}$

$$\frac{}{|\{i \in I_{a,b}\}|} \qquad \frac{}{|\{i \in I_{a,b}\}|}$$

where $I_{a,b} = \{i \mid i \in \{i, \ldots, m\}, P(y^{(i)} = 1 | x^{(i)}; \theta) \in (a,b)\}$

and $|S|$ denotes the size of the set S.

a) * $\theta \in \mathbb{R}^{n+1}$ be the maximum likelihood parameters learned after training a logistic regression

* $x_0^{(i)} = 1$

Problem: Prove that the above property holds true for the described logistic regression model over the range $(a,b) = (0,1)$

θ are maximum likelihood parameters

$$\Rightarrow \frac{\partial l_\theta}{\partial \theta_j} = \frac{\partial \sum_{i=1}^{m} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)})) \right)}{\partial \theta_j}$$

$$\Rightarrow \quad 0 = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - h(x^{(i)})) x_j^{(i)}$$

We know that $x_0^{(i)} = 1$

$$\Rightarrow \text{When } j=0 \quad \frac{1}{m} \sum_{i=1}^{m} y^{(i)} = \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)})$$

b) Problem

* Does well-calibrated model have perfect accuracy?
* Does perfect accuracy make a model well-calibrated?

Let's prove them wrong with a counterexample

- Assumptions: * $a = 0,6$; $b = 0,7$
  * all examples are positive

$$\Rightarrow \sum_{i \in I_{a,b}} \frac{\mathbb{1}\{y^{(i)} = 1\}}{|\{i \in I_{a,b}\}|} = 1$$

$$\Rightarrow \sum_{i \in I_{a,b}} \frac{P(y^{(i)} = 1 \mid x^{(i)}, \theta)}{|\{i \in I_{c,b}\}|} \leq 0.7$$

Eventhough the model is perfectly accurate, it is not well calibrated.

- Assumptions : * $a = 0,6$, $b = 0.7$

  * Well calibrated model

$\Rightarrow$ $P(y^{(i)} = 1 | x^{(i)}; \theta) \geq 0,5 \quad \forall i \in I_{a,b}$

$\Rightarrow$ All predictions are positive (1)

But Since the model is well calibrated : $\frac{\sum_{i \in I_{a,b}} \mathbb{1}\{y^{(i)} = 1\}}{|\{i \in I_{a,b}\}|} \leq 0.7$

$\Rightarrow$ Not all true labels are positive (2)

(1) + (2) $\Rightarrow$ Well calibrated models don't necessarily have perfect accuracy.

c) What effect does L2 regularization in logistic regression have on model calibration

$$l(\theta) = \sum_{i=0}^{m} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right) - \frac{\lambda}{2} \|\theta\|^2$$

With MLE parameters:

$$\Rightarrow \frac{\partial l(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - h(x^{(i)})) x_j - \lambda \theta_j = 0$$

when $j = 0$

$$\Rightarrow \frac{\partial l(\theta)}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) - \lambda \theta_0 = 0$$

$$\Rightarrow \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) + \lambda \theta_0 = \frac{1}{m} \sum_{i=1}^{m} y^{(i)}$$

$\Rightarrow$ The model is not well-calibrated.

3 Bayesian Interpretation of Regularization

a) * $\theta_{map} = \arg\max_{\theta} p(\theta | x, y)$

* $\theta_{mle} = \arg\max_{\theta} p(y | x, \theta)$

* We assume $p(\theta) = p(\theta | x)$

Problem: Show that $\theta_{map} = \arg\max_{\theta} p(y | x, \theta) p(\theta)$

$$\Rightarrow \theta_{map} = \arg\max_{\theta} \frac{p(y, x, \theta)}{p(x, y)} p(\theta)$$

$$= \arg\max_{\theta} \frac{p(y | x, \theta)}{p(x, y)} p(\theta, x)$$

$$= \arg\max_{\theta} p(y | x, \theta) \cdot \frac{p(\theta | x) p(x)}{p(x) \cdot p(y | x)}$$

$$= \arg\max_{\theta} \frac{p(y | x, \theta)}{p(y | x)} p(\theta)$$

$$= \arg\max_{\theta} p(y | x, \theta) p(\theta)$$

b) * $\theta \sim \mathcal{N}(0, \eta^2 I)$

* $\theta_{MAP} = \arg\max_{\theta} p(y|x,\theta)\, p(\theta)$

Problem: Show that

$$\theta_{MAP} = \arg\min_{\theta} -\log p(y|x,\theta) + \lambda \|\theta\|_2^2$$

and find the value of λ?

$$\theta_{MAP} = \arg\max_{\theta} p(y|x,\theta)\, p(\theta)$$

$$= \arg\min_{\theta} -\log\left(p(y|x,\theta)\, p(\theta)\right)$$

↑ log is a strictly increasing function

$$= \arg\min_{\theta} -\log\left(p(y|x,\theta)\right) - \log\left(p(\theta)\right)$$

* $\arg\min_{\theta} -\log(p(\theta)) = \arg\min_{\theta} -\log(c) - \log\left(\exp\left(-\frac{1}{2}\theta^T \Sigma^{-1} \theta\right)\right)$

$$= \arg\min_{\theta} \frac{1}{2}\, \theta^T \frac{1}{\eta^2} I\, \theta$$

$$= \arg\min_{\theta} \frac{1}{2\eta^2} \cdot \|\theta\|_2^2$$

$$\Rightarrow \theta_{MAP} = \arg\min_{\theta} -\log p(y|x,\theta) + \frac{1}{2\eta^2}\|\theta\|_2^2$$

$$\Rightarrow \lambda = \frac{1}{2\eta^2}$$

c) * Linear regression: $y = \theta^T x + \varepsilon$ (2)

* $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ (1)

* $\theta \sim \mathcal{N}(0, \eta^2 I)$ (3)

* $X = \begin{bmatrix} -x^{(1)}- \\ -x^{(2)}- \\ \vdots \\ -x^{(m)}- \end{bmatrix}$; $\vec{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix}$

Problem: Come up with a closed form expression for θ_{MAP}.

* (1) + (2) + (3) $\Rightarrow p(y^{(i)}|x^{(i)},\theta) = \mathcal{N}(x^T\theta, \sigma^2)$

* $p(y|x,\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta)$

$$\Rightarrow \arg\max_{\theta} p(y|x,\theta) = \arg\min -\log\left(\sum_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta)\right)$$

$$= \arg\min -\sum_{i=1}^{m} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{y^{(i)} - x^{(i)T}\theta}{\sigma}\right)^2$$

$$= \arg\min_{\theta} \frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta)$$

$$\Rightarrow \theta_{MAP} = \arg\min_{\theta} \frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta) + \underbrace{\frac{1}{2\eta^2}\|\theta\|_2^2}$$

Question (b)

$$* \quad \frac{d}{d\theta}\left(\frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta) + \frac{1}{2\eta^2}\|\theta\|_2^2\right)$$

$$= \frac{1}{2\sigma^2}\left(2X^TX\theta - 2X^TY\right) + \frac{1}{2\eta^2} \cdot 2\theta$$

$$= \left(X^TX\theta - X^TY\right)\frac{1}{\sigma^2} + \frac{1}{\eta^2}\theta$$

Setting $\frac{d}{d\theta}$ to 0

$$\Rightarrow \left(\frac{X^TX}{\sigma^2} - \frac{I}{\eta^2}\right)\theta = X^TY \cdot \frac{1}{\sigma^2}$$

$$\theta_{MAP} = \left(\frac{X^TX}{\sigma^2} - \frac{I}{\eta^2}\right)^{-1} X^TY \frac{1}{\sigma^2}$$

d.) * $y = x^T\theta + \varepsilon$

* $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

* $\theta \sim \mathcal{L}(0, bI)$

* $f_{\mathcal{L}}(z|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|z-\mu|}{b}\right)$

* $\theta_{MAP} = \underset{\theta}{\operatorname{argmin}} -\log(p(y|x,\theta)) - \log(p(\theta))$

Problem: Show that θ_{MAP} is equivalent to linear regression with L_1 regularization whose loss is specified as

$$J(\theta) = \|X\theta - \vec{y}\|_2^2 + \gamma\|\theta\|_1$$

$$\log(p(\theta)) = \cdots - \log\left(\frac{1}{2b}\right) - \frac{\|\theta\|_1}{b}$$

$$\Rightarrow \theta_{MAP} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta) - \log\left(\frac{1}{2b}\right) + \frac{\|\theta\|_1}{b}$$

$$= \underset{\theta}{\operatorname{argmin}} \|X\theta - \vec{y}\|_2^2 + \frac{\|\theta\|_1}{b}$$

$$\Rightarrow \gamma = \frac{1}{b}$$

4) Constructing Kernels

* K_1, K_2 are kernels over $\mathbb{R}^n \times \mathbb{R}^n$

* $a \in \mathbb{R}^+$

* $f : \mathbb{R}^n \to \mathbb{R}$

* $\phi : \mathbb{R}^n \to \mathbb{R}^d$

* K_3 kernel over $\mathbb{R}^d \times \mathbb{R}^d$

* $p(x)$ a polynomial over x with positive coefficients

Problem: For each K below, state whether it is a kernel or not.

a) $K(x,z) = K_1(x,z) + K_2(x,z)$

* $K_1 + K_2$ is a symmetric matrice cause K_1 and K_2 are symmetric

* $z^T(K_1 + K_2)z = z^T K_1 z + z^T K_2 z \geq 0$

$\Rightarrow$ K is a Kernel.

b) $K(x,z) = K_1(x,z) - K_2(x,z)$

If $K_2 = a K_1$

$\Rightarrow z^T K z = z^T K_1 z - a z^T K_1 z < 0$

K is not necessarily a kernel

c) $K(x,z) = a K_1(x,z)$

* $a K_1$ is symetric
* $z^T K z = a z^T K z \geq 0$

$\Rightarrow$ K is a kernel

d) $K(x,z) = -a K_1(x,z)$

$z^T K z = -a z^T K z < 0$

K is not a kernel

e) $K(x,z) = K_1(x,z)\,K_2(x,z)$

$\Rightarrow K_{ij} = \phi_1(x^{(i)})^T \phi_1(x^{(j)})\, \phi_2(x^{(i)})^T \phi(x^{(j)})$

$\Rightarrow z^T K z = \sum_i \sum_j z_i \phi_1(x^{(i)})^T \phi_1(x^{(j)}) \phi_2(x^{(i)})^T \phi_2(x^{(j)}) z_j$

$= \sum_i \sum_j \sum_k \sum_l z_i \phi_{1k}(x^{(i)}) \phi_{1k}(x^{(j)}) \phi_{2l}(x^{(i)}) \phi_{2l}(x^{(j)}) z_j$

$= \sum_k \sum_l \left(\sum_i z_i \phi_{1k}(x^{(i)}) \phi_{2l}(x^{(i)}) \right) \left(\sum_j z_j \phi_{1k}(x^{(j)}) \phi_{2k}(x^{(j)}) \right)$

$= \sum_k \sum_l \left(\sum_i z_i \phi_{1k}(x^{(i)}) \phi_{2l}(x^{(i)}) \right)^2 \geq 0$

$\Rightarrow K$ is a kernel.

f) $K(x,z) = f(x) f(z)$ where $f : \mathbb{R}^n \mapsto \mathbb{R}$

$\Rightarrow z^T K z = \sum_i \sum_j z_i f(x^{(i)}) f(x^{(j)}) z_j$

$= \left(\sum_i z_i f(x^{(i)}) \right) \left(\sum_j z_j f(x^{(j)}) \right) \geq 0$

$\Rightarrow K$ is a kernel.

g) The inputs x, z and don't have an effect on the validity of kernel K_3

h) $K(x,z) = p(K_1(x,z))$

where $p(x)$ is a polynomial with positive coefficients

$$\Rightarrow p(x) = \sum_{i=0}^{D} \alpha_i x^i$$

$$\Rightarrow p(K_1(x,z)) = \sum_{i=0}^{D} \alpha_i K_1(x,z)^i$$

a)
c)
e)
$\Rightarrow$ K is a valid kernel

5) Kernelizing the Perceptron

* $y \in \{0, 1\}$

* $h_\theta(x) = g(\theta^T x)$ where $g(z) = 1$ if $z \geq 0$, -1 otherwise

* Update Rule:

$$\theta^{(i+1)} := \theta^{(i)} + \alpha\left(y^{(i+1)} - h_{\theta^{(i)}}(x^{(i+1)})\right)x^{(i+1)}$$

* $\theta^{(0)} = \vec{0}$

* $K(x, z) = \phi(x)^T \phi(z)$ where ϕ is a high-dimensional feature mapping

i) According to Representer Theorem

$$\theta^{(i)} = \sum_{j=0}^{i} \alpha_j y_j \phi(x^{(j)}) \qquad y_j \in \{-1, 1\}$$

$$\text{or} = \sum_{j=0}^{i} \alpha_j \phi(x^{(j)}) \qquad \alpha_j \in \mathbb{R}$$

$$\theta^{(0)} = \vec{0}$$

(ii) How to calculate $h_{\theta^{(i)}}(x^{(i+1)})$ efficiently?

$$h_{\theta^{(i)}}(x^{(i+1)}) = g(\theta^{(i)T}\phi(x^{(i+1)}))$$

$$= \text{sign}\left(\sum_{j=1}^{i} \alpha_j \cdot \phi(x^{(j)})^T \phi(x^{(i+1)})\right)$$

$$= \text{sign}\left(\sum_{j=1}^{i} \alpha_j \cdot K(x^{(j)}, x^{(i+1)})\right)$$

(iii) Modify the update rule mentioned above to use the feature mapping

$$\theta^{(i+1)} := \theta^{(i)} + \alpha(y^{(i+1)} - h_{\theta^{(i)}}(x^{(i+1)}))x^{(i+1)}$$

$$\Rightarrow \theta^{(i+1)} := \sum_{j=1}^{i} \alpha_j \phi(x^{(j)}) + \alpha\left(y^{(i+1)} - \text{sign}\left(\sum_{j=1}^{i} \alpha_j \cdot K(x^{(j)}, x^{(i+1)})\right)\right)\phi(x^{(i+1)})$$

c) The dot kernel performs badly, cause it doesn't use feature mapping and the data is not linearly seperable