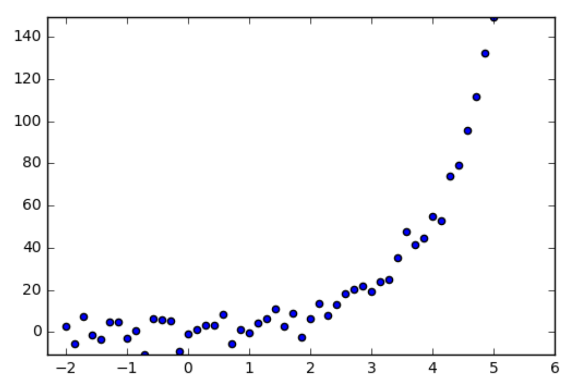
5. Linear Regression and Regularization

In this question, we will investigate the fitting of linear regression.

5. (a)

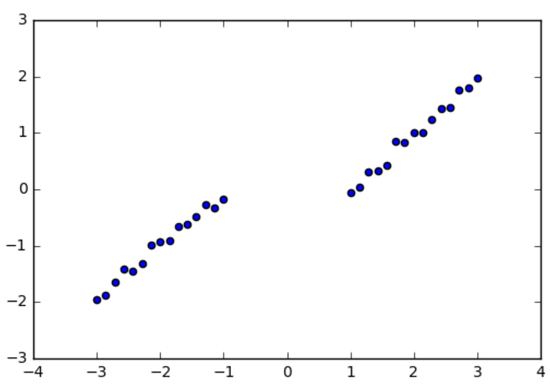
1/2 points (graded)

For each of the datasets below, provide a simple feature mapping ϕ such that the transformed data $(\phi(x^{(i)}), y^{(i)})$ would be well modeled by linear regression.



Which feature mapping ϕ is appropriate for the above model?

- (selected) $\exp(x)$ ✔
- $\log(x)$
- x^2
- $\sqrt{x}$



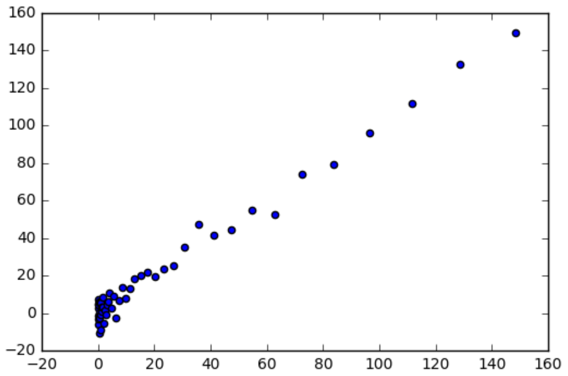
Which feature mapping ϕ is appropriate for the above model? 这里我理解他妈错了x轴是什么，x轴就是原来的x，y是被mapping过的x
要把x map到和y能有linear关系

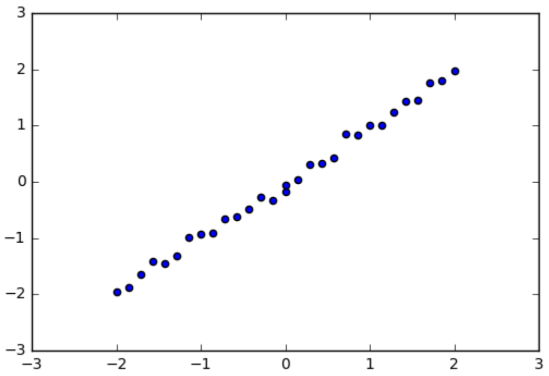
- $\phi(x) = x + \text{sign}(x)$
- $\phi(x) = x - \text{sign}(x)$ ✔

- $\phi(x) = x \cdot \mathrm{sign}(x)$
- $\phi(x) = x/\mathrm{sign}(x)$ ✖

Solution:

- In both figures the data seem to follow a non-linear pattern so they would not be fit well by a linear model.
- We can, however, use a non-linear transformation $\phi(x)$ so that, in the new feature space, a linear model produces a good fit.
- In the 1st plot, the data seem to roughly follow $y = e^x$, so an exponential transformation, $\phi(x) = e^x$, would yield $(\phi(x^{(i)}), y^{(i)})$ that could be fit well by linear regression.
- In the 2nd plot, the observations appear to be generated by the discontinuous function $y = x - \mathrm{sign}(x)$ (where $\mathrm{sign}(x) = x/|x|$), so if we let $\phi(x) = x - \mathrm{sign}(x)$, an observation $y^{(i)}$ should be more easily modeled by a linear function of $\phi(x^{(i)})$, which will be found by linear regression.
- The results of the transformations are plotted below.





Submit You have used 2 of 2 attempts

Answers are displayed within the problem

5. (b)

2.0/2 points (graded)

Consider fitting a ℓ_2-regularized linear regression model to data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $x^{(t)}, y^{(t)} \in \mathbb{R}$ are scalar values for each $t = 1, \ldots, n$. To fit the parameters of this model, one solves

$$\min_{\theta \in \mathbb{R},\, \theta_0 \in \mathbb{R}} L(\theta, \theta_0)$$

where

$$L(\theta, \theta_0) = \sum_{t=1}^{n} \left(y^{(t)} - \theta x^{(t)} - \theta_0\right)^2 + \lambda \theta^2$$

Here $\lambda \geq 0$ is a pre-specified fixed constant, so your solutions below should be expressed as functions of λ and the data. This model is typically referred to as **ridge regression** .

Write down an expression for the gradient of the above objective function in terms of θ.

Important: If needed, please enter $\sum_{t=1}^{n}(\ldots)$ as a function `sum_t(...)`, including the parentheses. Enter $x^{(t)}$ and $y^{(t)}$ as `x^{t}` and `y^{t}`, respectively.

$\frac{\partial L}{\partial \theta} =$ sum_t(-2*x^{t}*(y^{t} - theta*x^{t} - theta_0)) + 2*lambda ✔

Answer: 2*lambda*theta - 2*sum_t((y^{t} - theta*x^{t} - theta_0)*x^{t})

Write down an expression for the gradient of the above objective function in terms of θ_0.

$\frac{\partial L}{\partial \theta_0} =$ sum_t(-2*(y^{t} - theta*x^{t} - theta_0)) ✔ **Answer:** -2*sum_t(y^{t} - theta*x^{t} - theta_0)

STANDARD NOTATION

Solution:

- The gradient is a two-dimensional vector $\nabla L = \left[\frac{\partial L}{\partial \theta_0}, \frac{\partial L}{\partial \theta}\right]$, where
- $\frac{\partial L}{\partial \theta_0} = -2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)} - \theta_0\right)$
- $\frac{\partial L}{\partial \theta} = 2\lambda\theta - 2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)} - \theta_0\right)x^{(t)}$

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

5. (c)

2.0/2 points (graded)
Find the closed form expression for θ_0 and θ which solves the ridge regression minimization above.

Assume θ is fixed, write down an expression for the optimal $\hat{\theta}_0$ in terms of $\theta, x^{(t)}, y^{(t)}, n$.

Important: If needed, please enter $\sum_{t=1}^{n}(\ldots)$ as a function `sum_t(...)`, including the parentheses. Enter $x^{(t)}$ and $y^{(t)}$ as `x^{t}` and `y^{t}`, respectively.

$\hat{\theta}_0 =$ (sum_t(y^{t} - theta*x^{t}))/n ✔ **Answer:** 1/n * sum_t(y^{t} - theta*x^{t})

Write down an expression for the optimal $\hat{\theta}$. To simplify your expression, use $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x^{(t)}$. Your answer should be in terms of $x^{(t)}, y^{(t)}, \lambda$ and $\bar{x}$ **only**.

Important: If needed, please enter $\sum_{t=1}^{n}(\ldots)$ as a function `sum_t(...)`, including the parentheses. Enter $x^{(t)}$ and $y^{(t)}$ as `x^{t}` and `y^{t}`, respectively. Enter $\bar{x}$ as `barx`.

$\hat{\theta} =$ (sum_t(x^{t}*y^{t})-barx*sum_t(y^{t})) /(sum_t((x^{t})^2)+ ✔

Answer: (sum_t((x^{t} - barx)*y^{t})) / (lambda + sum_t(x^{t} * (x^{t} - barx)))

Now after the optimal $\hat{\theta}$ is obtained, you can use it to compute the optimal $\hat{\theta}_0$

Solution:

To find the θ, θ_0 which minimize L, we note that because this objective function is convex, any point where $\nabla L(\theta_0, \theta) = 0$ is a global minimum. Thus, we set the gradient equal to zero and solve for θ, θ_0 to find the minimizers:

$$\frac{\partial}{\partial \theta_0} = -2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)} - \theta_0\right) = -2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)}\right) + 2\sum_{t=1}^{n}\theta_0 = 0$$

$$\implies \quad -2n\theta_0 = -2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)}\right) \quad \implies \quad \theta_0 = \frac{1}{n}\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)}\right)$$

$$\frac{\partial}{\partial \theta} = 2\lambda\theta - 2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)} - \theta_0\right)x^{(t)}$$

$$= 2\lambda\theta - 2\sum_{t=1}^{n}\left(y^{(t)} - \theta x^{(t)} - \left[\frac{1}{n}\sum_{s=1}^{n}\left(y^{(s)} - \theta x^{(s)}\right)\right]\right)\cdot x^{(t)} = 0$$

$$\implies \quad \lambda\theta - \sum_{t=1}^{n} x^{(t)}y^{(t)} + \theta\sum_{t=1}^{n} {x^{(t)}}^2 + \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\left(y^{(s)} - \theta x^{(s)}\right)x^{(t)} = 0$$

$$\implies \lambda\theta - \sum_{t=1}^{n} x^{(t)} y^{(t)} + \theta \sum_{t=1}^{n} x^{(t)^2} + \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} y^{(s)} x^{(t)} - \frac{1}{n} \theta \sum_{t=1}^{n} \sum_{s=1}^{n} x^{(s)} x^{(t)} = 0$$

$$\implies \hat{\theta} = \frac{\sum_{t=1}^{n} x^{(t)} y^{(t)} - \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} y^{(s)} x^{(t)}}{\lambda + \sum_{t=1}^{n} x^{(t)^2} - \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} x^{(s)} x^{(t)}} \quad \text{is the value of } \theta \text{ which minimizes } L(\theta_0, \theta).$$

Note that if we define $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x^{(t)}$, then we can rewrite the above expression in a nicer form:

$$\hat{\theta} = \frac{\sum_{t=1}^{n} \left(x^{(t)} - \bar{x}\right) y^{(t)}}{\lambda + \sum_{t=1}^{n} x^{(t)} \left(x^{(t)} - \bar{x}\right)}$$

In other words, adding an unpenalized bias is equivalent to training on a centered dataset.

Finally, we can plug this value of $\hat{\theta}$ back into expression $\hat{\theta}_0 = \frac{1}{n} \sum_{t=1}^{n} \left(y^{(t)} - \theta x^{(t)}\right)$ to find the corresponding $\hat{\theta}_0$ which together with $\hat{\theta}$ minimizes L.

Submit You have used 2 of 5 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 2 Nonlinear Classification, Linear regression, Collaborative Filtering (2 weeks):Homework 3 / 5. Linear Regression and Regularization