XCS299i Problema Set #1

JUAN RICARDO PEDRAZA ESCOBAR

Machine Learning

Stanford Center for Professional Development

July,2021

1.a

We will take into account the following equalities in order to develop the demonstration:

(i)
$$E\ [\ Y;\eta] = \int yp\ \ (y;\eta)\ \ dy$$

(ii)
$$\int p\ (y;\eta)\ \ dy = 1$$

And also looking at the statement,
$$\frac{\partial}{\partial\eta}\int p\ (y;\eta)\ \ dy = \int \frac{\partial}{\partial\eta} p\ (y;\eta)\ \ dy$$

Considering that the first derivative of the logarithmic function $a(\eta)$ with respect to the natural parameters η is the expectation of a distribution of the exponential family,
$$\frac{\partial}{\partial\eta}\ p(y;\eta) = \ p(y;\eta)\left(y - \frac{\partial}{\partial\eta}\ a(\eta)\right)$$

$$\int \frac{\partial}{\partial\eta}\ p(y;\eta)\ \ dy = \int p(y;\eta)\left(y - \frac{\partial}{\partial\eta}\ a(\eta)\right)\ \ dy$$

Using the statement hint,
$$\frac{\partial}{\partial\eta}\int\ p(y;\eta)\ \ dy = \int yp(y;\eta)dy - \int p(y;\eta)\frac{\partial}{\partial\eta}\ a(\eta)\ \ dy$$

Using the equality marked with (i) on the right side and the equality marked with (ii) on the left side of the equation, we obtained,
$$\frac{\partial}{\partial\eta}\ 1 = \ E\ [\ Y;\eta] - \frac{\partial}{\partial\eta}\ a(\eta)\int\ p(y;\eta)\ \ dy$$

We use the equality marked with (ii) again in the left said of the equation,

$$0 = \ E\ [\ Y;\eta] - \frac{\partial}{\partial\eta}\ a(\eta)$$

We finally obtain,

$$\frac{\partial}{\partial\eta}\ a(\eta) = E\ [\ Y;\eta]$$

1.b

Using the result of the previous exercise (1a), we start with,

$$\frac{\partial}{\partial \eta} a(\eta) = E\,[Y; \eta]$$

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = \frac{\partial}{\partial \eta} E\,[Y; \eta]$$

Using the equality marked (i) in the previous exercise (1a) in the right side,

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = \frac{\partial}{\partial \eta} \int yp\ (y; \eta)\ dy$$

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = \int p(y; \eta) \left(y - \frac{\partial}{\partial \eta} a(\eta) \right)\ dy$$

$$= \int y^2 p(y; \eta) dy - \frac{\partial}{\partial \eta} a(\eta) \int yp(y; \eta)\ \ dy$$

Using again the equality marked (i) in the previous exercise (1a) in the right side,

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = E\,[Y^2; \eta] - E\,[Y; \eta]^2$$

We finally obtain,

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = Var\,(Y; \eta)$$

1.c

The negative log likelihood loss function has the following form,

$$\ell\,(\theta) = -\log[p(y;\eta)]$$

$$\ell\,(\theta) = -\log\left[\left(b(y)\;\; exp(\eta y - a(\eta))\right]\right.$$

$$\ell\,(\theta) = a(\eta)\; - \eta y\; + C$$

Writing de function in terms of θ ,

$$\ell\,(\theta) = a(\,\theta^T x) - \theta^T x\, y\; + C$$

The function is derived with respect to θ in order to find the gradient of the loss function,

$$\nabla_\theta \ell\,(\theta) = \frac{\partial}{\partial \eta} a(\eta)\;\; \nabla_\theta \eta - yx$$

$$\nabla_\theta \ell\,(\theta) = \frac{\partial}{\partial \eta} a(\eta) x - yx$$

In order to find de Hessian of the loss function we find the second derivative with respect to θ,

$$\nabla_\theta^{\,2}\; \ell\,(\theta) = \;\nabla_\theta\left(\frac{\partial}{\partial \eta} a(\eta) x - yx\right)$$

$$\nabla_\theta^{\,2}\; \ell\,(\theta) = \;\nabla_\theta\left(\frac{\partial}{\partial \eta} a(\eta) x - yx\right)$$

$$\nabla_\theta^{\,2}\; \ell\,(\theta) = x\;\frac{\partial}{\partial \eta}\left(\frac{\partial}{\partial \eta} a(\eta)\right)\nabla_\theta \eta$$

$$\nabla_\theta^{\,2}\; \ell\,(\theta) = \;\frac{\partial^2}{\partial \eta^2}\; a(\eta)\; xx^T$$

Using the equivalent that we find in part b in the right side,

$$\nabla_\theta^{\,2}\; \ell\,(\theta) = Var\;(Y;\eta)\;\; xx^T$$

- Var(Y ; η) is positive for any value of θ.
- The Hessian of the negative log-likelihood of the Generalized Linear Model is PSD, and consequently convex.

2.a

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{N} \left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)^2$$

Differentiating this objective, we get:

$$\nabla_\theta J(\theta) = \sum_{i=1}^{N} \left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)\left(\hat{x}^{(i)}\right)$$

$$= \lambda \sum_{i=1}^{N} \left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)\left(\hat{x}^{(i)}\right)$$

The gradient descent update rule is;

$$\theta := \theta - \lambda \nabla_\theta J(\theta)$$

which reduces here to:

$$\theta := \theta - \lambda \sum_{i=1}^{N} \left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)\left(\hat{x}^{(i)}\right)$$

where λ denotes the learning rate.

2.d

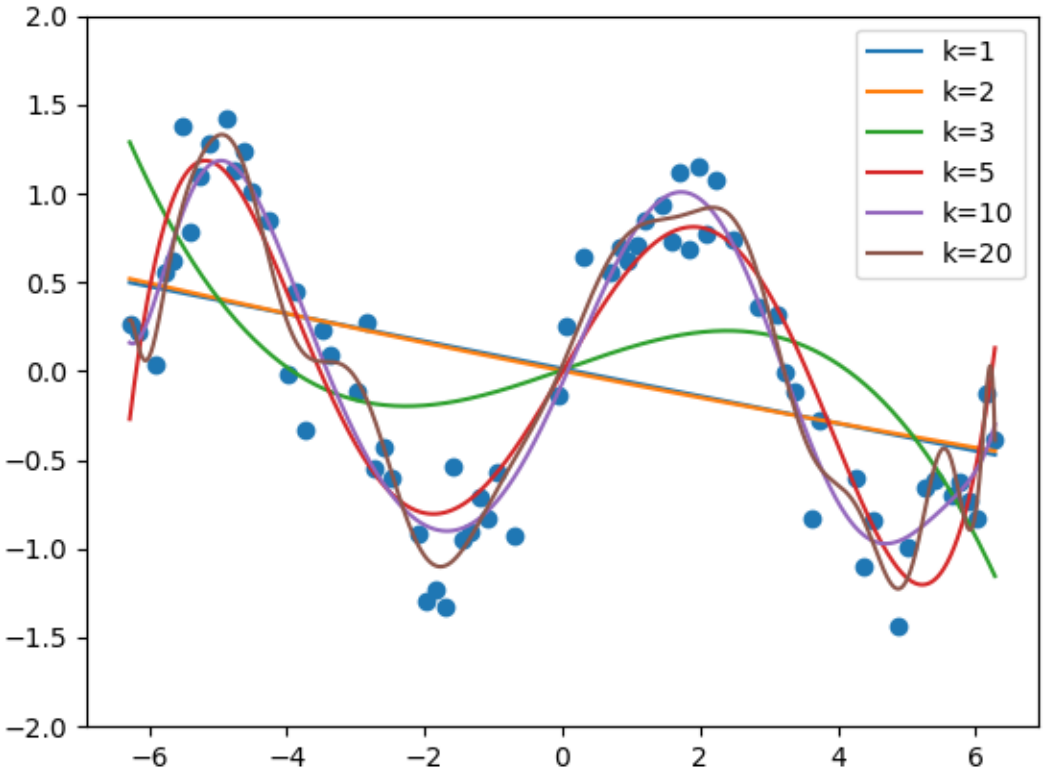

Figure 1: Polynomial regression with kernel sizes 1,2,3,5,10 and 20.

- Polynomials with values of k=1 and k=2 which can be seen as straight lines, in no way follows the curvature of training examples.
- The polynomial whit value of k=3, due to its degree, presents curvatures. However, these are not sufficient to fit the curvature of the training examples in a correct way.
- Polynomials with values of k = 5 and k = 10, due to their high degree, are a better fit to the training samples.
- The polynomial whit value of k=20, due to very high degree, shows abnormal curvatures because the model starts to overfit to the dispersion in the training set. It shows that the models with very high degree polynomials can be numerically non-stable.

2.f

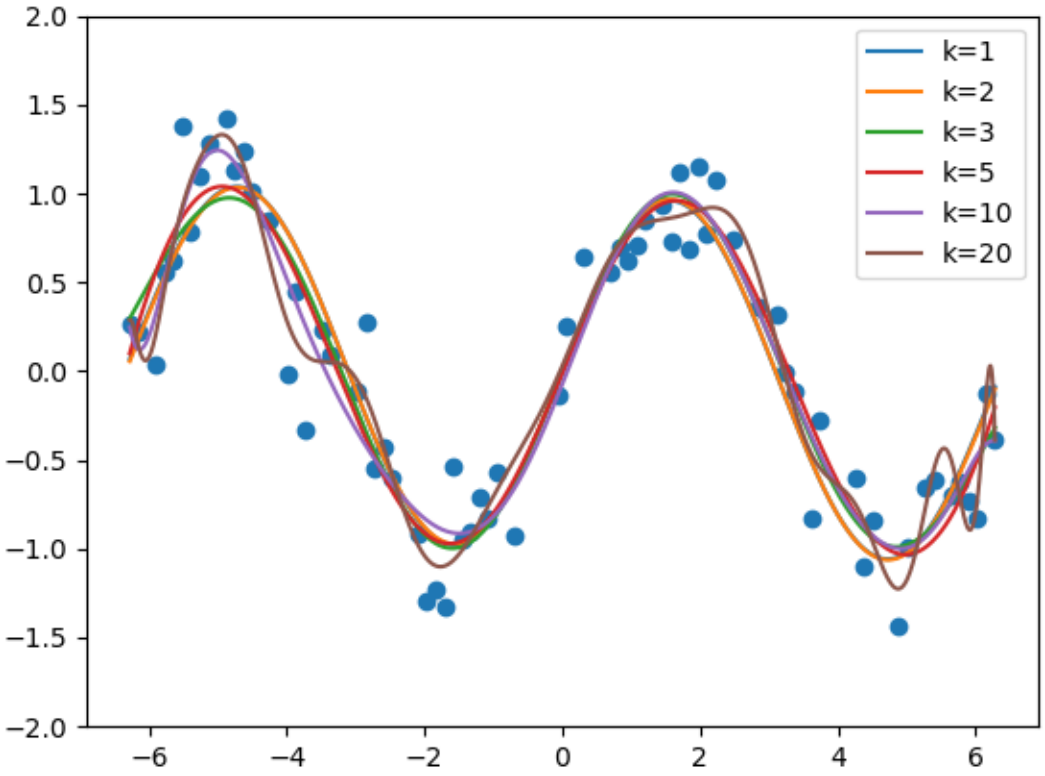

Figure 2: Polynomial regression with other features with kernel sizes 1,2,3,5,10 and 20

- Looking at the figure, the sin(x) feature allows the model to better fit the training samples, even with low degree polynomial terms (k=1, k=2, k=3). However, for very high degree polynomials (k=20) the numerical instability remains, showing abnormal curvatures in different parts of it.

2.h

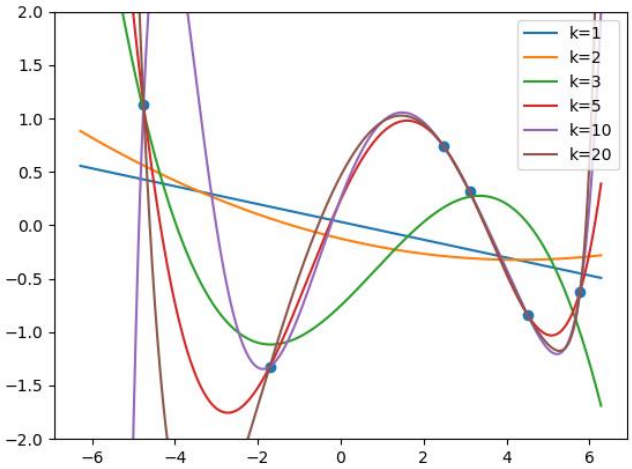

Figure 3: Polynomial regression with kernel sizes 1,2,3,5,10 and 20 on small dataset

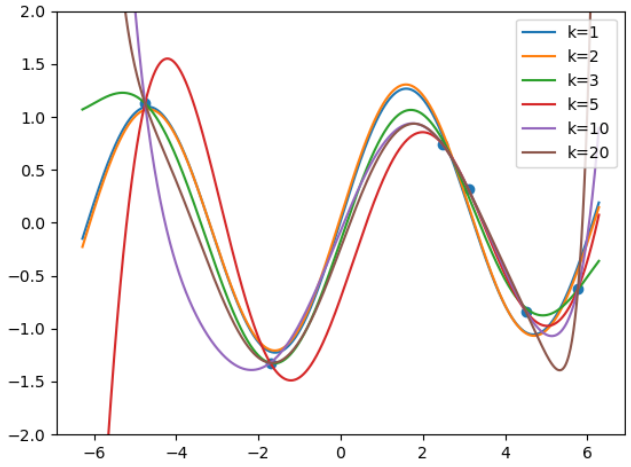

Figure 4: Regression with other polynomial and sinusoidal features with kernel sizes 1,2,3,5,10 and 20 on small dataset

- We see that the polynomials of the highest degree pass through all the points in the training examples when the number of parameters is greater than the number of training examples (small data set), but the curves appear to have a poor fit. Also, high degree polynomials are still having the problem of numerical instability, even with sin (x) feature and a small data set.