Homework 6

Due April 5 at 11 pm

Yves Greatti - yg390

1. (Gradient descent and ridge regression) In this problem we study the iterations of gradient descent applied to the ridge-regression cost function

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \left|\left|y - X^T\beta\right|\right|_2^2 + \frac{\lambda}{2} \left|\left|\beta\right|\right|_2^2, \tag{1}$$

where $X \in \mathbb{R}^{p \times n}$ is a fixed feature matrix and $y \in \mathbb{R}^n$ is a response vector. (The factor of 1/2 is just there to make calculations a bit cleaner.)

(a) Derive a closed form expression for the value of the estimated coefficient $\beta^{(k)}$ at the kth iteration of gradient descent initialized at the origin in terms of the SVD of X when the step size is constant.

The gradient-descent updates are:

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \ \alpha_k \nabla_\beta f(\beta^{(k)}) \\
&= \beta^{(k)} - \alpha_k (XX^T\beta^{(k)} - Xy + \lambda\beta^{(k)}) \\
&= ((1 - \lambda\alpha_k)I - \alpha_k XX^T)\beta^{(k)} + \alpha_k Xy \\
&= \left((1 - \alpha\lambda)I - \alpha XX^T\right)^{k+1}\beta^{(0)} + \alpha \sum_{i=0}^{k} \left((1 - \alpha\lambda)I - \alpha XX^T\right)^i Xy \\
&= \alpha \sum_{i=0}^{k} \left((1 - \alpha\lambda)I - \alpha XX^T\right)^i Xy
\end{aligned}$$

since the step size $\alpha_k = \alpha$ is constant and $\beta^{(0)}$ is the zero vector (initialization at the origin). Let the svd of $X = USV^T$ then

$$\beta^{(k+1)} = \alpha \sum_{i=0}^{k} \left((1 - \alpha\lambda)I - \alpha US^2U^T\right)^i USV^Ty$$

Assuming $p \leq n$ and X is full rank, $UU^T = U^TU = I$ and we have:

$$\begin{aligned}
\beta^{(k+1)} &= \alpha \sum_{i=0}^{k} \left((1-\alpha\lambda)UU^T - \alpha US^2U^T\right)^i USV^Ty \\
&= \alpha U \sum_{i=0}^{k} \left((1-\alpha\lambda)I - \alpha S^2\right)^i SV^Ty \\
&= \alpha U \text{diag}_{j=1}^{p} \sum_{i=0}^{k} \left(1 - \alpha(s_j^2+\lambda)\right)^i SV^Ty \\
&= U \text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^{k+1} s_j}{s_j^2+\lambda} \, V^Ty
\end{aligned}$$

(b) Under what condition on the step size does gradient descent converge to the ridge-regression coefficient estimate as $k \to \infty$?
If step size α is small enough: $0 < \alpha < \frac{2}{\lambda+s_1^2} \leq \frac{2}{\lambda+s_j^2} \to |1-\alpha(s_j^2+\lambda)| < 1$ then $\lim_{k\to\infty}(1-\alpha(s_j^2+\lambda))^k = 0, j = 1, \ldots, p$, gradient descent converges to:

$$\begin{aligned}
\lim_{k\to\infty} \beta^{(k)} &= U \; \text{diag}_{j=1}^{p} \left(\frac{s_j}{s_j^2+\lambda}\right) V^Ty \\
&= U(S^2+\lambda I)^{-1}S^2V^Ty
\end{aligned}$$

which are the ridge-regression coefficient estimates.

(c) Assume the following additive model for the data:

$$\tilde{y}_{\text{train}} := X^T\beta_{\text{true}} + \tilde{z}_{\text{train}}, \tag{2}$$

where $\tilde{z}_{\text{train}}$ is modeled as an n-dimensional iid Gaussian vector with zero mean and variance σ^2. What is the distribution of the estimated coefficient $\tilde{\beta}^{(k)}$ at the kth iteration of gradient descent initialized at the origin?
Using the expression of the estimated coefficients from part a, we now have:

$$\begin{aligned}
\tilde{\beta}^{(k)} &= U \text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j}{s_j^2+\lambda} \, V^T(X^T\beta_{\text{true}} + \tilde{z}_{\text{train}}) \\
&= U \text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j}{s_j^2+\lambda} \, V^T(VSU^T\beta_{\text{true}} + \tilde{z}_{\text{train}}) \\
&= U \text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j^2}{s_j^2+\lambda} \, U^T\beta_{\text{true}} + U \text{diag}_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j}{s_j^2+\lambda} \, V^T\tilde{z}_{\text{train}}
\end{aligned}$$

Using theorem 8,6 from the notes on PCA, then the estimated coefficient $\tilde{\beta}^{(k)}$ at the kth iteration of gradient descent initialized at the origin have is a Gaussian random vector with mean:

$$\beta_{\text{GD}} = \sum_{j=1}^{p} \frac{1-(1-\alpha(s_j^2+\lambda))^k s_j^2}{s_j^2+\lambda} \langle u_j, \beta_{\text{true}} \rangle \, u_j$$

and covariance matrix

$$\Sigma_{\text{GD}} = \sigma^2 U \text{diag}_{j=1}^{p} \frac{(1-(1-\alpha(s_j^2+\lambda))^k)^2 s_j^2}{(s_j^2+\lambda)^2} U^T$$

(d) Complete the script *RR_GD_landscape.py* in order to verify your answer to the previous question. Report the figures generated by the script.

2. (Climate modeling) In this problem we model temperature trends using a linear regression model. The file `t_data.csv` contains the maximum temperature measured each month in Oxford from 1853-2014. We will use the first 150 years of data (the first $150 \cdot 12$ data points) as a training set, and the remaining 12 years as a test set.

   In order to fit the evolution of the temperature over the years, we fit the following model

   $$y[t] = a + bt + c\cos(2\pi t/T) + d\sin(2\pi t/T) \tag{3}$$

   where $a, b, c, d \in \mathbb{R}$, $y[t]$ denotes the maximum temperature in Celsius during month t of the dataset (with t starting from 0 and ending at $162 \cdot 12 - 1$).

   (a) What is the number of parameters in your model and how many data points do you have to fit the model? Are you worried about overfitting?

   (b) Fit the model using least squares on the training set to find the coefficients for values of T equal to 1,2,...,20. Which of these models provides a better fit? Explain why this is the case. In the remaining question we will fix T to the value T^* that provides a better fit.

   (c) Produce two plots comparing the actual maximum temperatures with the ones predicted by your model for $T := T^*$; one for the training set and one for the test set.

   (d) Fit the modified model

   $$y[t] = a + bt + d\sin(2\pi t/T^*) \tag{4}$$

   and plot the fit to the training data as in the previous question. Explain why it is better to also include a cosine term in the model.

   (e) Provide an intuitive interpretation of the coefficients a, b, c and d, and the corresponding features. According to your model, are temperatures rising in Oxford? By how much?

3. (Sines and cosines) Let $x : [-1/2, 1/2) \to \mathbb{R}$ be a real-valued square-integrable function defined on the interval $[-1/2, 1/2)$, i.e. $x \in L_2[-1/2, 1/2)$. The Fourier series coefficients of x, are given by

   $$\hat{x}[k] := \langle x, \phi_k \rangle = \int_{-1/2}^{1/2} x(t) \exp\left(-i2\pi kt\right) \, \mathrm{d}t, \quad k \in \mathbb{Z}, \tag{5}$$

   and the corresponding Fourier series of order k_c equals

   $$\mathcal{F}_{k_c}\{x\}(t) = \sum_{k=-k_c}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right). \tag{6}$$

   As we will discuss in class, this is a representation of x in a basis of complex exponentials. In this problem we show that for real signals the Fourier series is equivalent to a representation in terms of cosine and sine functions.

(a) Prove that $\hat{x}[k] = \overline{\hat{x}[-k]}$ for all $k \in \mathbb{Z}$. [Hint: What is $\overline{e^{it}}$?]

$$\begin{aligned}
\overline{\hat{x}[-k]} &= \overline{\int_{-1/2}^{1/2} x(t) \exp\left(i2\pi kt\right)\, \mathrm{d}t} \\
&= \int_{-1/2}^{1/2} \overline{x(t)\exp\left(i2\pi kt\right)}\, \mathrm{d}t \\
&= \int_{-1/2}^{1/2} x(t) \exp\left(-i2\pi kt\right)\, \mathrm{d}t \\
&= \hat{x}[k]
\end{aligned}$$

(b) Show that the Fourier series of x of order k_c can be written as

$$\mathcal{F}_{k_c}\{x\}(t) = a_0 + \sum_{k=1}^{k_c} a_k \cos(2\pi kt) + b_k \sin(2\pi kt),$$

for some $a_0, \dots, a_k, b_1, \dots, b_k \in \mathbb{R}$. [Hint: Group terms in $\mathcal{F}_{k_c}\{x\}(t)$ corresponding to $\pm k$ and use previous part. What is the real part of zw for $z, w \in \mathbb{C}$?]

$$\begin{aligned}
\mathcal{F}_{k_c}\{x\}(t) &= \sum_{k=-k_c}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right) \\
&= \hat{x}[0] + \sum_{k=-k_c}^{-1} \hat{x}[k] \exp\left(i2\pi kt\right) + \sum_{k=1}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right) \\
&= \hat{x}[0] + \sum_{k=k_c}^{1} \hat{x}[-k] \exp\left(-i2\pi kt\right) + \sum_{k=1}^{k_c} \hat{x}[k] \exp\left(i2\pi kt\right) \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} (\hat{x}[-k] \exp\left(-i2\pi kt\right) + \hat{x}[k] \exp\left(i2\pi kt\right)) \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} (\overline{\hat{x}[k]} \exp\left(-i2\pi kt\right) + \hat{x}[k] \exp\left(i2\pi kt\right)) \quad \text{using part a}
\end{aligned}$$

Given two complex numbers $z = a+ib$ and $w = c+id$, we have $zw = ac-bd+i(ad+bc)$ and $\overline{z}\,\overline{w} = ac - bd - i(ad + bc)$ giving that $zw + \overline{zw} = 2(ac - bd)$. Let $z_k = \hat{x}[k]$ and

$w_k = \exp(i2\pi kt)$ thus

$$\begin{aligned}
\mathcal{F}_{k_c}\{x\}(t) &= \hat{x}[0] + \sum_{k=1}^{k_c} 2(\mathrm{Re}\,(\hat{x}[k])\cos 2\pi kt - \mathrm{Im}\,(\hat{x}[k])\sin 2\pi kt) \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} (2\,\mathrm{Re}\,(\hat{x}[k]))\cos 2\pi kt + (-2\,\mathrm{Im}\,(\hat{x}[k]))\sin 2\pi kt \\
&= \hat{x}[0] + \sum_{k=1}^{k_c} a_k\cos(2\pi kt) + b_k\sin(2\pi kt)
\end{aligned}$$

(c) Give expressions for the coefficients a_k, b_k for $k \geq 1$ from the previous part as real integrals. Interpret them in terms of inner products.

From the definition

$$\begin{aligned}
\hat{x}[k] = \langle x, \phi_k\rangle &= \int_{-1/2}^{1/2} x(t)\exp(-i2\pi kt)\ \mathrm{d}t \text{ for } k \geq 1 \\
&= \int_{-1/2}^{1/2} x(t)(\cos(2\pi kt) + i\sin(2\pi kt))\,\mathrm{d}t \\
&= \int_{-1/2}^{1/2} x(t)\cos(2\pi kt)\,\mathrm{d}t + i\int_{-1/2}^{1/2} x(t)\sin(2\pi kt)\,\mathrm{d}t \\
&= \mathrm{Re}\,(\hat{x}[k]) + i\,\mathrm{Im}\,(\hat{x}[k])
\end{aligned}$$

thus

$$a_k = 2\int_{-1/2}^{1/2} x(t)\cos(2\pi kt)\,\mathrm{d}t = 2\langle x, \cos(2\pi kt)\rangle = 2\langle x, \mathrm{Re}\,(\phi_k)\rangle$$

and

$$b_k = 2\int_{-1/2}^{1/2} x(t)\sin(2\pi kt)\,\mathrm{d}t = 2\langle x, \sin(2\pi kt)\rangle = 2\langle x, \mathrm{Im}\,(\phi_k)\rangle$$

.

(d) Suppose $x(t) = \cos(2\pi(t+\phi))$ for some fixed $\phi \in \mathbb{R}$. What are the Fourier coefficients of x?

(e) Suppose that f is also even (i.e., $x(-t) = x(t)$). Prove that the Fourier coefficients are all real (i.e., that $\hat{x}[k] \in \mathbb{R}$ for all $k \in \mathbb{Z}$).

Using part a

$$
\begin{aligned}
\overline{\hat{x}[k]} &= \hat{x}[-k] \\
&= \int_{-1/2}^{1/2} x(t) \exp\left(i2\pi kt\right)\, \mathrm{d}t \\
&= \int_{1/2}^{-1/2} x(-u) \exp\left(-i2\pi ku\right) \left(-\,\mathrm{d}u\right) \text{ by change of variable } u = -t \\
&= \int_{-1/2}^{1/2} x(t) \exp\left(-i2\pi kt\right)\, \mathrm{d}t \quad \text{since x is even} \\
&= \hat{x}[k]
\end{aligned}
$$

$\overline{\hat{x}[k]} = \hat{x}[k]$, $\hat{x}[k] \in \mathbb{R}$ for all $k \in \mathbb{Z}$.