SUGGESTED EXERCISES: MODERN REGRESSION

JEAN-LUC BOUCHOT

1. Simple linear regression

Homework 1. Prove the following statement

Proposition 1.1 (Prove at home)**.** *Under the conditions that the LINE assumptions are valid, the following partition of errors hold*

$$SSTO = SSR + SSE.$$

Homework 2 (Optional). Try to prove the following relation

Proposition 1.2. *Pearson's correlation coefficient can be computed as*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Homework 3. Getting some practice with t distributions. You are trying to find the t multiplier in the definition of a confidence interval for a (under certain assumptions on your data) t distributed variable.

(1) Recall the expression of a confidence interval for a t-distributed random variable.
(2) What is the t multiplier if you have 15 samples and ask for a 95% confidence interval?
(3) What is the t multiplier if you have 26 samples and ask for a 95% confidence interval?
(4) What is the t multiplier if you have 15 samples and ask for a 91% confidence interval?

Homework 4. Suppose that you produce dragon fruits and you've noticed that you produce 55% of the red kind and 45% of the white kind. You have in front of you 100 dragon fruits 53 of which are red. Can you conclude that the sample is representative of your production?

2. Multiple linear regression: more dimensions

3. Matrices: what you need to know

Homework 5. What happens to inverses if the matrix A is not square? Let $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 2/3 & -1/3 & 1/3 \\ -1/3 & 2/3 & 1/3 \end{pmatrix}$. Show that BA is some identity (which one?) but AB clearly is not. This shows that we may have left inverses which are not right inverses.

Homework 6. Let A be any square matrix. Show that the matrix $C = \frac{A - A^T}{2}$ is antisymmetric.

Homework 7. Prove the following using Laplace formula.

Proposition 3.1. *Let $A \in \mathbb{R}^{3\times 3}$. We may use Sarrus' rule to compute a 3×3 determinant:*

$$\begin{aligned}\det(A) = {} & A_{1,1}A_{2,2}A_{3,3} + A_{1,2}A_{2,3}A_{3,1} + A_{1,3}A_{2,1}A_{3,2} \\ & - A_{1,3}A_{2,2}A_{3,1} - A_{1,2}A_{2,1}A_{3,3} - A_{1,1}A_{2,3}A_{3,2}.\end{aligned}$$

Homework 8. Show that the equivalent of Sarrus rule (sum of positive diagonals minus sum of negative diagonals) doesn't work for dimension 4.

Homework 9. Show the following result on inverses of 2 dimensional matrices

Date: May 19, 2021.

Proposition 3.2. *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{R}^{2\times 2}$*. If* A *is invertible, its inverse is given by*

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Homework 10. Let $D, U, L \in \mathbb{R}^{n\times n}$ where D is a diagonal matrix, U is an upper triangular matrix (i.e. $U_{i,j} = 0$ if $i > j$) and L is a lower triangular matrix (i.e. $L_{i,j} = 0$ if $j > i$). Find $\det(D), \det(L), \det(U)$.

Homework 11. Determine whether the following vectors are orthogonal or not

- $\mathbf{x} = (6,1,4)^T$ and $\mathbf{y} = (2,0,-3)^T$.
- $\mathbf{x} = (0,0,-1)^T$ and $\mathbf{y} = (1,1,1)^T$.
- $\mathbf{x} = (0,0,-1)^T$ and $\mathbf{y} = (-1,-1,0)^T$.
- $\mathbf{x} = (a,0,0)^T$ and $\mathbf{y} = (0,0,b)^T$ for some numbers a and b.

Homework 12. Let $\mathbf{x} = (-7,-18,\alpha)^T$ and $\mathbf{y} = (0,-10,-7)^T$. Find the value(s) of α for which $\mathbf{x}$ and $\mathbf{y}$ are perpendicular.

Homework 13. Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Show that A is positive definite.

Homework 14. Let L be a one-dimensional subspace (i.e. a line) in an n dimensional space $\mathbb{R}^n$. For any vector $x \in \mathbb{R}^n$, find the projection of this x onto the line L.

Homework 15. Let $f(x,y) = 3x^2 - 2xy + y^2$ and define $x(u,v) = 3u + 2v$ and $y(u,v) = 4u - v$. Compute ∇f using the chain rule. Verify your result with a direct calculation (i.e. replacing the expression of x and y and computing the gradient directly.)

Homework 16. Let $A \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$. Find the gradient of the function

$$f(x) = \|Ax - b\|^2.$$

What are the critical points of the function f?

Homework 17. Let $f(x,y,z) = xy + yz + zx$. Using the formula for bilinear forms, compute the gradient of f.

Homework 18. Let $f(x) = \|Ax-b\|_2^2 + \lambda\|x\|_2^2$ for some matrix $A \in \mathbb{R}^{m\times d}$, some vector $b \in \mathbb{R}^m$ and a constant $\lambda \geq 0$. Find the critical points of f.

Homework 19. Let $f_a : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be a function defined as $f(x) = xa^T$ for a given vector $a \in \mathbb{R}^d$. Find the Jacobian of f_a.

Homework 20. Let $f : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be defined as $f(x) = xx^T$. Compute the Jacobian of f.

Define further $g : \mathbb{R}^d \to \mathbb{R}$ be defined, for a matrix $A \in \mathbb{R}^{d\times d}$, as

$$g(x) = \|A - xx^T\|^2$$

where the norm is defined for any matrix $B \in \mathbb{R}^{d\times d}$ as $\|B\|^2 = \sum_{i,j} B_{i,j}^2$. Compute the gradient of g.

(Note: assuming we are trying to minimize the error, the resulting matrix is called the rank 1 approximation of A – A concept very important in recommender systems such as used by Netflix and Amazon)

4. Multiple linear regression

Homework 21. Let y be the response variable and assume we have access to two predictors x_1 and x_2: $x = (x_1, x_2)$. Define

$$f_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1^2 x_2 + \beta_6 x_2^3.$$

Does f_β define a MLR model?

Homework 22. Let $y_i = \mathbf{x}\beta + \varepsilon_i$ where all the ε_i's are all independently identically distributed according to a Normal with 0 mean and variance σ^2. Show that the least squares estimator is unbiased.

Show that the matrix of covariances of $\widehat{\beta}$ is given by

$$\mathrm{Cov}(\widehat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

5. Regularization and model selection

Homework 23. Let $A \in \mathbb{R}^{(d+1)\times(d+1)}$ be the data matrix containing the $(d+1)$ polynomial features of $d+1$ different sampling points $(x_i)_{0\leq i\leq d}$.

Show that, if $x_i \neq x_j$ for all $i \neq j$, then A is invertible. Use this to prove Lagrange's interpolation theorem, i.e., for any $d+1$ points, there exists an interpolating polynomial of degree d.

Homework 24. Let $\mathcal{D} = \{(x_i, y_i), 0 \leq i \leq n\}$ be a set of datapoints such that $x_i \neq x_j$ for all $i \neq j$. Define

$$\ell_j(x) = \prod_{\substack{i=0\\ i\neq j}}^{n} \frac{x - x_i}{x_j - x_i}$$

and

$$L_{\mathcal{D}}(x) = \sum_{j=0}^{n} y_j \ell_j(x).$$

Show the followings:

1. $\ell_j(x_i) = 0$ if $i \neq j$.
2. $\ell_j(x_j) = 1$.
3. $L_{\mathcal{D}}(x_i) = y_i$ for all $0 \leq i \leq n$.
4. Prove Lagrange Interpolation theorem.

Homework 25. For any n distinct points, there exists a unique interpolating polynomial.

True or False? Justify.

Homework 26 (This is an empirical exercise for gaining experience with the tools; it's not exam style). With the help of a computer, test the following:

1. Choose a (multiple) linear regression model based on polynomial features of your choice (for instance $y = 2x^2 + x - 3$). We will call this $f_\theta(x)$ with θ the regression coefficients.
2. Generate n sampling location uniformly in $[-1, 1]$, with $n > d+1$, where d is the highest power in your polynomial features.
3. Generate the target variable using the sampling locations and the regression formula you have chosen.
4. Use a least squares estimator to recover the regression coefficients. How do they compare with the truth? Is it expected?
5. Without changing the true model, vary the degree of the polynomial (from something smaller than the truth to something greater than the truth) that you are trying to fit and compute the regression coefficient. What happens? Is it expected?
6. Consider now some noise in the measurements (i.e. generate your sampling points as $y_i = f_\theta(x_i) + \varepsilon_i$ with ε_i a normally distributed random variable with mean 0 and a given – small – variance). Repeat the two previous experiments. Comment on what happens.

From now on, we will try to understand the impact of noise and how to detect overfitting. Fix your true measurement model $y = f_\theta(x)$ and some noise standard deviation $\sigma > 0$ as well as the (known) sampling locations x_i for $1 \leq i \leq n$ (make sure n is larger than the max degree in your model. Say if $d = 2$ is your highest degree, use about $n = 30$ data points). Repeat the following operations at least 50 times:

1. Generate some noisy measurements $y_i = f_\theta(x_i) + \varepsilon_i$.
2. Using the true degree d, compute the linear regression with polynomial features of degree d.
3. Save the results.

Once done, compute the empirical expectation (= the average) of the regression coefficients and compare with the true coefficients. What is the variance of the coefficients?

Now consider two polynomial degrees which you want to test if they are a good fit to the data: $d_1 < d < d_2$. (for instance, consider $d_1 = 1$ and $d_2 = 7$, if you had $d = 2$). Redo the same as you just did. What can be said about the expectations and variances of the coefficients?

Homework 27. Let $A \in \mathbb{R}^{n\times D}$ with $D \leq n$. Show the following:

1. If $\lambda \neq 0$ is an eigenvalue of A^TA, then it is an eigenvalue of AA^T.
2. Verify that the condition $\lambda \neq 0$ is an important one.

(3) If λ is an eigenvalue of A^TA then $\lambda \geq 0$.
(4) Show that $\left(A^TA + \lambda I\right)$ is non singular.

Homework 28. Let

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

(1) Compute the eigenvalues of A^TA (you might notice that $\lambda = 1$ is an eigenvalue).
(2) Compute the eigenvalues of $A^TA + \lambda I$ for $\lambda > 0$.
(3) Compare the sum of the squared eigenvalues of A^TA and the sum of its squared eigenvalues.

Homework 29. Show that the ridge regression can be expressed as an ordinarily least squares which uses modified versions of the data matrix X and the observations $\mathbf{y}$.

Homework 30. Assume the dataset $\mathcal{D} = \{(1.4, 0), (1.4, -2), (0.8, 0), (0.4, 2)\}$. Find the parameter $\lambda > 0$ such that the regression coefficient of a linear regression with linear features is

$$\widehat{\beta}(\lambda) = \left(1, -\frac{1}{8}\right)^T.$$

Homework 31. Let $\widehat{\beta}(\lambda)$ be the estimated parameters of a ridge regression with parameter $\lambda > 0$. Show that the bias variance reads

$$MSE(\widehat{\beta}(\lambda)) = \mathbb{E}[\|\beta - \widehat{\beta}(\lambda)\|^2] = \text{trace}\left(\text{Var}(\widehat{\beta})\right) + \|\text{bias}\|^2.$$

Homework 32. Let $X \in \mathbb{R}^{n\times d}$ be our data matrix (containing n observations/samples/individuals each of dimension d). Define $\widehat{\beta}(\lambda)$ the ridge regression estimators obtained from the sample values $\mathbf{y} \in \mathbb{R}^n$ and with parameter $\lambda \geq 0$.

(1) Using a singular value decomposition, show that there exist two orthogonal matrices U and V such that

$$\widehat{\beta}(\lambda) = VD_\lambda U^T\mathbf{y},$$

where

$$D_\lambda = \text{diag}(\sigma_i/(\sigma_i^2 + \lambda))$$

with σ_i the (nonzero) singular values of X.
(2) How can you express the predicted values $\widehat{\mathbf{y}}$?

Homework 33. This exercise shows the importance of pre processing data before doing any learning on them.

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with ε_i identically and independently sampled from a normal with 0 mean and variance σ^2. The independent variables are given by $x = (x_i)_{i=1}^8 = (-2, -1, -1, -1, 0, 1, 2, 2)$ and the targets are given by $\mathbf{y} = (y_1, y_2, \cdots, y_8)^T = (35, 40, 36, 38, 40, 43, 45, 43)^T$.

(1) Find the ridge regression estimator for the data above for a general value of λ.
(2) Evaluate the fit, i.e. $\|\mathbf{y} - \widehat{\mathbf{y}}(\lambda)\|_2^2$ for $\lambda = 10$. Would you judge the fit as good? If not, what is the most striking feature that you find unsatisfactory?
(3) Now center the data (i.e. subtract the mean) and target values, and denote them by $\widetilde{x}_i$ and $\widetilde{y}_i$ and evaluate the ridge estimator $\widetilde{\mathbf{y}}$ using $\widetilde{x}$. Check the fit for $\lambda = 4$ and $\lambda = 10$.

6. General gradient descent approaches

Homework 34. Let us look at descent methods for ridge regression. Consider the following optimization problem:

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_2^2,$$

for a given full rank matrix $A \in \mathbb{R}^{n\times d}$ and a regularization coefficient $\lambda > 0$.

- Define the gradient descent steps.
- Define the Newton descent steps.
- Suggest what could be a possible option for the quasi-Newton step.

Homework 35. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as

$$f(x) = \frac{2}{3}|x|^{3/2}.$$

Define a step size $\gamma > 0$.

(1) Show that f admits a unique minimizer.
(2) Define $x^* = \left(\frac{\gamma}{2}\right)^2$. Detail the sequence of iterates given by the gradient descent starting with $x^0 = x^*$.
(3) Consider now $x^0 \in (0, x^*)$. What can be said about the sequence of iterates?
(4) Conclude with regards to the use of gradient descent without care.

Homework 36. Let $\Omega = \{x \in \mathbb{R}^d : x_i \geq 0, \forall 1 \leq i \leq d\}$. Compute P_Ω, the projection operator onto Ω.

School of Mathematics and Statistics, Beijing Institute of Technology
Email address: jlbouchot@bit.edu.cn