NYU Paris **Spring 2019**

CSCI-UA 9473 Introduction to Machine Learning

Problem Set 2: Linear regression, regularization, model selection

1. Multivariable differential calculus and chain rule

Exercise 1. *Using the chain rule, compute the derivative of the function* $y = \log(\sin^2(x^3))$

Exercise 2 (source: Andrew Ng, CS 229)**.** *Consider the function* $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}$*. What is the gradient of this function? the Hessian* ($\nabla^2 f$)*? Same question with* $f(x) = g(\boldsymbol{a}^T\boldsymbol{x})$*.*

Exercise 3. *Find the partial derivatives* $\frac{\partial f}{\partial x}$ *and* $\frac{\partial f}{\partial y}$ *for the following functions*

- $f(x, y) = \cos(x^2 y) + y^5$
- $f(x, y) = e^{x^2+y^3}$
- $f(x, y) = \sqrt{1 - x^2 - y^3}$
- $f(x, y) = x \tan(y^2)$
- $f(x, y) = \frac{1}{xy}$

Exercise 4. *Consider a differentiable function* $f(x, y)$*. Give physical interpretation of the meaning of* $\frac{\partial f}{\partial x}(a, b)$ *and* $\frac{\partial f}{\partial y}$ *in terms of the graph of* $f(x, y)$*. I.e. what do the partial derivatives represent?*

Exercise 5. *Use the chain rule to compute the following derivatives*

- $\frac{dz}{dt}$ *for* $z = \sin(x^2 + y^3)$*,* $x = t^2 + 3$*,* $y = t^2$
- $\frac{dz}{dt}$ *for* $z = x^2 y$*,* $x = \cos(t)$ *and* $y = t^2 + 2$
- $\frac{\partial z}{\partial t}$ *and* $\frac{\partial z}{\partial s}$ *for* $z = x^3 y$ *and* $x = \cos(st)$ *and* $y = t^2 + s$
- $\frac{\partial f}{\partial x}$ *and* $\frac{\partial f}{\partial t}$ *for* $z = x^2 y^3$ *and* $x = st$ *and* $y = t^2 - s^2$

Exercise 6. *Compute all first and second order derivatives of the following functions*

- $f(x, y) = \frac{xy}{x^2+y^2}$
- $f(x, y) = 4x^3 + xy + 5$
- $f(x, y) = \sin(5x)\cos(3y)$
- $\log(\sqrt{x^2 + y^3})$

Exercise 7. *Compute all first and second order derivatives of the following functions*

- $f(x, y) = \frac{xy}{x^2+y^2}$
- $f(x, y) = 4x^3 + xy + 5$
- $f(x, y) = \sin(5x)\cos(3y)$
- $\log(\sqrt{x^2 + y^3})$

Exercise 8 (source: D. Guichard)**.** *Find and characterize all local extremas of the following functions*

- $f(x, y) = x^2 + 4y^2 - 2x + 8y - 1$
- $f(x, y) = x^2 - y^2 + 6x - 10y + 2$
- $f(x, y) = xy$

2. Linear algebra and matrix derivatives

Exercise 9. *Prove the following relations*

- $\mathrm{Tr}(\boldsymbol{AB}) = \mathrm{Tr}(\boldsymbol{BA})$
- $\frac{\partial \boldsymbol{x}^T \boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial x} = \boldsymbol{a}$
- $\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T$
- $\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{b}\boldsymbol{a}^T$
- $\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{a}^T$

Exercise 10. *Prove the following relations*

- $\frac{\partial \mathrm{Tr}(\boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{I}$
- $\frac{\partial \mathrm{Tr}(\boldsymbol{XA})}{\partial \boldsymbol{X}} = \boldsymbol{A}^T$
- $\frac{\partial}{\partial \boldsymbol{X}} \mathrm{Tr}(\boldsymbol{AXB}) = \boldsymbol{A}^T \boldsymbol{B}^T$
- $\frac{\partial}{\partial \boldsymbol{X}} \mathrm{Tr}(\boldsymbol{X}^2) = 2\boldsymbol{X}^T$
- $\frac{\partial}{\partial \boldsymbol{X}} \mathrm{Tr}(\boldsymbol{X}^2 \boldsymbol{B}) = (\boldsymbol{XB} + \boldsymbol{BX})^T$

Exercise 11. *Given a vector space $\mathcal{V}$, a norm $n(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is a function $\mathcal{V} \mapsto [0, +\infty)$ which satisfies the following properties*

- $n(\boldsymbol{x} + \boldsymbol{y}) \leq n(\boldsymbol{x}) + n(\boldsymbol{y})$
- $n(\alpha \boldsymbol{x}) = |\alpha| n(\boldsymbol{x})$
- $n(\boldsymbol{x}) = 0 \iff \boldsymbol{x} = 0$

Consider the norms $n(\boldsymbol{x}) = \sum_{i=1}^N |x_i|$ (ℓ_1 norm) and $\|\boldsymbol{x}\|_\infty = \sup_i |x_i|$ (ℓ_∞ norm). Show that those norms satisfy the properties above.

3. Linear regression and regularization

Exercise 12 (source: CB)**.** *We have seen that the linear regression model finds the line that minimizes the sum of orthogonal distances of the prototypes $\boldsymbol{x}_\mu$ to the line. For a particular data point $(x', y') \in \mathbb{R}^2$, show that the point on a line $y = a + bx$ that is the closest when we measure distance orthogonally is given by $(\hat{x}', \hat{y}')$ where $\hat{x}'$ and $\hat{y}'$ are defined as*

$$\hat{x}' = \frac{by' + x' - ab}{1 + b^2}, \quad \hat{y}' = a + \frac{b}{1 + b^2}(by' + x' - ab) \tag{1}$$

Hint: use Pythagorean Theorem.

Exercise 13 (source: CB)**.** *Consider the one dimensional regression problem for a set of pairs (x_i, y_i),*

$$\min_{a,b} \sum_{i=1}^N (y_i - (a + bx_i))^2 \tag{2}$$

The easiest way to find the minimum of this regression function is to compute the partial derivatives and set those derivatives to 0. However, this minimum can also be found as follows.

- *We now use (x_i, y_i) to denote a set of pairs and $(\hat{x}_i, \hat{y}_i)$ to denote the set of points generated from the regression model. The total least squares problem is to minimize the objective*

$$\sum_{i=1}^{n}((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) \tag{3}$$

- *One can expand the objective (3) as*

$$\sum_{i=1}^{n}((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) \tag{4}$$

$$= \sum_{i=1}^{n}\left(\frac{b^2}{(1+b^2)^2}[y_i - (a + bx_i)]^2 + \frac{1}{(1+b^2)^2}[y_i - (a + bx_i)]^2\right) \tag{5}$$

$$= \frac{1}{1+b^2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 \tag{6}$$

For a fixed b, the term in front of the sum is constant. The minimizing choice for a is thus given by $a = \overline{y} - b\overline{x}$ (why?)

- *If we substitute this expression back into (6), we get that the total least squares solution for b is the one that minimizes*

$$\frac{1}{1+b^2}\sum_{i=1}^{n}\left((y_i - \overline{y}) - b(x_i - \overline{x})\right)^2 \tag{7}$$

Now introduce the expressions

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{8}$$

The objective (7) can read as

$$\frac{1}{1+b^2}\left[S_{yy} - 2bS_{xy} + b^2S_{xx}\right] \tag{9}$$

Show that the extrema of

$$f(b) = \frac{1}{1+b^2}\left[S_{yy} - 2bS_{xy} + b^2S_{xx}\right] \tag{10}$$

are given by

$$b = \frac{-(S_{xx} - S_{yy} \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy^2}})}{2S_{xy}} \tag{11}$$

Then show that the "+" solution gives the minimum of $f(b)$.

Exercise 14. *We now want to consider a linear regression model in higher dimension. We stack the prototypes $\boldsymbol{x}_\mu$ in a matrix $\boldsymbol{X}$ (each $\boldsymbol{x}_\mu$ is now a row of $\boldsymbol{X}$) and write the regression problem as*

$$RSS(\beta) = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{12}$$

Show that the minimum of this function is given by $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

You can make use of the following results from matrix algebra

$$\frac{\partial(\boldsymbol{b}^T\boldsymbol{a})}{\partial \boldsymbol{a}} = \boldsymbol{b} \tag{13}$$

$$\frac{\partial(\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a})}{\partial \boldsymbol{a}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{a} \tag{14}$$

$$\frac{\partial}{\partial \boldsymbol{A}}\mathrm{tr}(\boldsymbol{B}\boldsymbol{A}) = \boldsymbol{B}^T \tag{15}$$

$$\frac{\partial}{\partial \boldsymbol{A}}\log(|\boldsymbol{A}|) = \boldsymbol{A}^{-T} = (\boldsymbol{A}^{-1})^T \tag{16}$$

$$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}) \tag{17}$$

Here the "trace" of a matrix $\boldsymbol{A}$ denotes the sum of the diagonal elements of $\boldsymbol{A}$, $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{N} \boldsymbol{A}_{ii}$

Exercise 15 (sources: HTF, Murphy)**.** *The statistical intuition behind linear regression is a model of the form*

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{\beta}^T\boldsymbol{x}, \sigma^2) \tag{18}$$

where $\mathcal{N}(y|\boldsymbol{\beta}^T\boldsymbol{x}, \sigma^2)$ is used to denote the Gaussian distribution with mean $\mu =$. If we put oursleves in a Bayesian framework, and we decide to consider additional priors, we get different regularization terms. An example of such regularized linear models, the ridge regression model reads as

$$\hat{\beta}_{ridge} = \underset{\beta}{\mathrm{argmin}} \left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{19}$$

Show that the ridge regression estimate is the mean (and mode) of the posterior distribution under a Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \tau\boldsymbol{I})$, and Gaussian sampling model, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ^2

Exercise 16. *Show that the solution to the ridge regression problem*

$$\underset{\boldsymbol{w}}{\mathrm{argmin}} \frac{1}{N} \sum_{i=1}^{N}(y_i - (w_0 + \boldsymbol{w}^T\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{w}\|_2^2 \tag{20}$$

can read as

$$\hat{\boldsymbol{w}}_{ridge} = (\lambda\boldsymbol{I}_D + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{21}$$

Exercise 17 (source: HTF)**.** *Show that the ridge regression estimate $\hat{\beta}_{RR}$ (i.e the solution* (19)*) can be obtained by ordinary least squares regression on an augmented dataset. We augment the centered matrix $\boldsymbol{X}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$ and augment $\boldsymbol{y}$ with p zeros (By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients towards* 0*).*

Exercise 18 (HTF)**.** *The "elastic net" formulation overcomes the limitations of the LASSO by combining the ℓ_2 and ℓ_1 penalties,*

$$\min_{\beta} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\left\{\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1\right\} \tag{22}$$

Show how one can turn this into a LASSO problem, using augmented versions of $\boldsymbol{X}$ and $\boldsymbol{y}$

Exercise 19 (source: HTF)**.** *Suppose that we run a Ridge regression with parameter λ on a single variable X and we get coefficient a. We now include an exact copy $X^* = X$ and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show that, in general, if m copies of a variable X_j are included in a ridge regression, their coefficients are all the same.*

Exercise 20 (source: HTF)**.** *The LASSO estimate can equivalently be written as the constrained problem*

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{23}$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{24}$$

Suppose that for w given t in (24)*, the fitted LASSO coefficient for the variable X_j is $\hat{\beta}_j = a$. Suppose that we augment the set of variables with an identical copy X_j^* of X_j, i.e. $X_j^* = X_j$. Characterize the effect of this exact colinearity by describing the set of solutions for $\hat{\beta}_j$ and $\hat{\beta}_j^*$ using the same value for t*

Exercise 21. *Show that the Ridge and LASSO estimators can be obtained as MAP estimators on a Gaussian likelihood with Gaussian (resp. Laplace) priors. The Laplace prior reads as*

$$P(\beta) = 2\lambda \exp(-\lambda|\beta|) \tag{25}$$