23 Problems

Aayush Bajaj

December 26, 2024

Contents

Rules

1. This year, I shall not advocate aversion from the internet.
2. If your work is unpleasant to read, and / or difficult to mark, I reserve the right to discard it.
3. The boxed numbers in the right margin are marks.
4. Deadline:*11:59PM*, 31st of January 2025.
5. Submission: LaTeX appraised, hand-written accepted. FILENAME MUST BE YOUR FULL NAME!

Sauce Code

Submit

§ Problems

§§ Appetisers

1. The Newton-Raphson method.

   (a) Estimate $\sqrt{2}$ using 2 iterations of the Newton-Raphson method. Let $x_0 = 1$ [2]
   Hint: The equation $f(x) = x^2 - 2$ may be helpful.

   (b) What does this method do, and how can it be used to find the optima of a function? [1]

Half mark each

   (c) List one advantage and one disadvantage for this method in practise. [1]
   Hint: Generalise to the multivariate matrix case.

2. Let n be the number of people in a group. What number must n be such that 2 people in this group share the same birthday with at least 50% probability? [2]

   (a) What number must n be such that 2 people share the same birthday with 100% probability? [1]

3. What kind of matrices possess an inverse? If the following matrices do, find them:

   (a) [1]

$$\begin{bmatrix} 5 & 4 & 3 \\ 3 & 2 & 1 \end{bmatrix}$$

   (b) [1]

$$\begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 6 \\ 4 & 6 & 8 \end{bmatrix}$$

   (c) [1]

$$\begin{bmatrix} 4 & 3 & 2 & 1 \\ 0 & 3 & 2 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

   Recall that $A|I \rightsquigarrow I|A^{-1}$

4. Let $\mathbf{x}^T = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T$ and $\mathbf{y}^T = \begin{bmatrix} -2 & 1 & 3 \end{bmatrix}^T$. Compute the *distance* between these two vectors by taking the **inner product** to be:

   (a) the familiar dot product: $\mathbf{x}^T\mathbf{y}$ [1]

   (b) $\mathbf{x}^T\mathbf{A}\mathbf{y}, \mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ [1]

   Hint: Recall that $\|x\| \coloneqq \sqrt{\langle x, x \rangle}$

5. Matrix Decompositions

(a) The trace of a matrix is equal to the _____ of the eigenvalues. [1]

(b) The determinant of a matrix is equal to the _____ of its eigenvalues. [1]

(c) Find the eigenvalues of this $\mathbb{R}^{2\times 2}$ matrix: [1]

$$\begin{bmatrix} 3 & 3 \\ 7 & -1 \end{bmatrix}$$

(d) *Qualitatively* describe the differences between eigendecompositions and Singular Value Decompositions[1] [1]

6. Given that a matrix $A \in \mathbb{R}^{n\times n}$ is *positive semi-definite*, denoted $A \succeq 0$, if $A = A^T$ and $x^T A x \geq 0,\ \forall x \in \mathbb{R}^n$. Show that $A = zz^T$ for an arbitrary $z \in \mathbb{R}^n$ [2]

7. Here are two ordinary, linear, autonomous and inhomogenous differential equations:

(a) Find the exact solution to: [2]

$$\frac{\mathrm{d}y}{\mathrm{d}x} + \cos(x)y = \cos(x), \quad y(\frac{\pi}{2}) = -1$$

(b) Find the general solution to: [2]

$$y'' + 14y' + 49y = e^t$$

8. List out the first four terms of the power series expansions of: [6]

(a) $\exp z$

(b) $\sin z$

(c) $\cos z$

(d) $\sinh z$

(e) $\cosh z$

(f) $\frac{1}{1-x}$

(g) $\frac{1}{1+x}$

9. Solve the following recurrence relation with initial conditions[2]: [3]

$$a_{n+1} = \frac{2a_n}{n+3}, \qquad a_0 = 1$$

10. Compute the following determinant — *efficiently.* [3]

$$\begin{vmatrix} 2 & 0 & 1 & 2 & 0 \\ 2 & -1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 \\ -2 & 0 & 2 & -1 & 2 \\ 2 & 0 & 0 & 1 & 1 \end{vmatrix}$$

[1] mathematics is sufficient, although not necessary here.

[2] be grateful for the ease of this recurrence; solving real ones in DE's are painful

11. Express the derivative of $\sigma(z) = \dfrac{1}{1+e^{-z}}$, $\sigma'(z)$, as a function of $\sigma(z)$. [2]

12. Describe the following Probability Distributions (in a sentence or two), and give an example experiment for each: [5]
    (a) Bernoulli
    (b) Binomial
    (c) Gaussian
    (d) Pareto
    (e) Geometric
    (f) Poisson
    (g) Uniform

13. "The principle point of proof is to compel belief" - Daniel Velleman. Keeping this in mind, explain to me either mathematically or otherwise, why [2]

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

§§ Main Course: Least Squares

The following section is devoted to deriving closed forms for the univariate and multivariate least squares model. We will then go on to see how to massage this linear model to fit obviously non-linear relationships (Figure 1, Image 3) and non-obvious non-linear relationships (Figure 2).

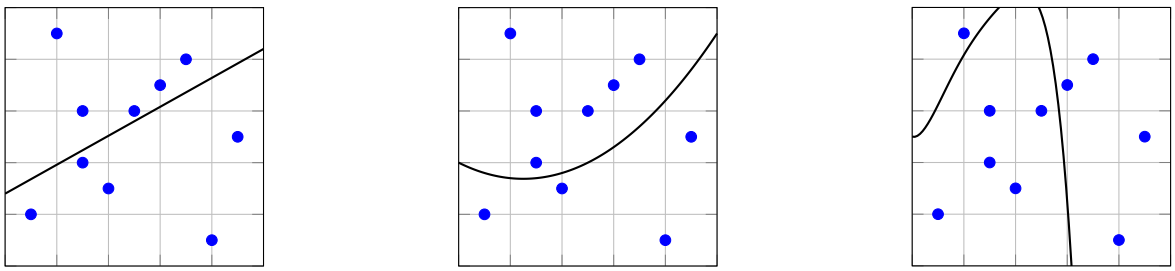

Figure 1: Linear Regression with increasingly polynomial features

§§§ Matrix Calculus

14. A *univariate* linear regression model is a model that tries to predict an outcome based on just a single independent variable: x. The model takes in data in the form of tuples and constructs a prediction (first plot, Figure 1) that we can use either to interpolate or extrapolate values of the dependent variable.

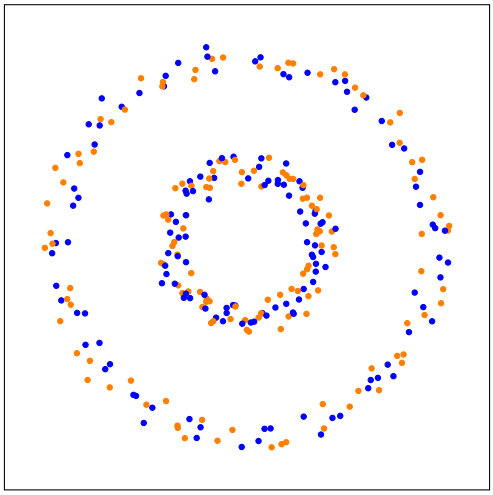

Figure 2: Non-Euclidean classification

The question then becomes, what kind of metric should we use to determine this line? In practice we construct a loss function $\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$. This takes each data point we were given, y_i and squares the difference between it and our linear model $y_i^* = w_0 + w_1 x_i$.[3]

Our loss function is then equivalently

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (w_0 + w_1 x_i))^2$$

(a) Obviously we want to minimise this loss function to achieve the best possible linear model[4]. We will see why this *Mean-Squared Error* (MSE) is a good choice in the following parts, but for now show that the w_0 which minimises this loss function is equal to $\overline{y} - w_1\overline{x}$. *Recall that* $\overline{a} = \frac{1}{n}\sum_{i=0}^{n} a_i$. [2]

(b) Repeat the steps by taking the partial derivative w.r.t w_1 and showing that the minimum occurs at $\dfrac{\overline{xy} - w_0\overline{x}}{\overline{x^2}}$. [2]

the hat above the y signifies *the* best estimate

(c) Solve the above weights simultaneously to rewrite the univariate regression line $\hat{y}(x) = w_0 + w_1 x$ in terms of the data only. [1]

15. We now extend this result to the multivariate case of p features[5] with n feature vectors[6]: $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$.

[3] the $*$ indicates *our* best estimate

[4] in the sense that the least squares estimate of the weight parameters will have the smallest variance amongst all linear unbiased estimates

[5] this is how many independent variables we have

[6] this is how many data points you have

So each i^{th} feature vector, x_i has p entries:

$$x_i = \begin{bmatrix} x_{i,0} \\ x_{i,1} \\ \vdots \\ x_{i,p-1} \end{bmatrix}$$

.

Our outputs (still y) will now be taking in richer feature vectors and applying weights $w_0, ..., w_{p-1}$ to each data point of each independent variable.

It makes sense to stack these feature vectors and construct a design matrix $X \in \mathbb{R}^{n \times p}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \vdots \\ \mathbf{x_n}^T \end{bmatrix} = \begin{bmatrix} x_{1,0} & x_{1,1} & \dots & x_{1,p-1} \\ x_{2,0} & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}$$

You should verify for yourself that the multivariate model now looks like:

$$\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_{p-1} x_{i,p-1} \tag{1}$$

And thus our previously manageable loss function transforms into

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_{p-1} x_{i,p-1})^2 \tag{3}$$

It should now be clear that taking partials with respect to each of the p weights is infeasible, and thus we must leverage matrix notation and calculus:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

where $\| \cdot \|_2^2$ denotes the L2 norm and has the property $\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x}$.

DO NOT FEAR SUMMATION EXPANSIONS!

(a) Begin by showing that $\nabla_{\mathbf{x}} \mathbf{b}^T \mathbf{x} = \mathbf{b}$, where $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$ [1]

(b) Next show that $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x}$, with $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x} \in \mathbb{R}^n$ [1]

(c) Finally, by leveraging the above rules, find $\nabla_w \mathcal{L}(\mathbf{w})$. [2]

(d) Set the above gradient to 0, and thus find the critical point. [1]

16. Whilst a critical point is a necessary condition for a minimum, it is not sufficient. Given that any critical point of a convex function *is* a minimum, show that the Hessian of $\mathcal{L}(\mathbf{w})$ is positive semidefinite. [1]
Recall the definition from Question 6.

17. Wielding this Hessian, apply the Newton-Raphson method to this problem to find out how many iterations it takes to find the optimal solution. Let $\mathbf{w} = w_0$. [2]

§§§ Statistics

In this section we will justify the use of the Mean-Squared Error that we befriended in the previous section. We will justify *mathematically* its existence as the best loss function under the assumption that the noise within our sampling was / is *normally distributed*.

18. Given a loss function parametrised on θ, $\mathcal{L}(\theta)$, we now want to choose this parameter that gives us the highest possible likelihood of observing the data. In other words, we wish to find the *maximum likelihood estimator* (MLE):

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$$

    (a) Assuming $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p), compute $\hat{p}_{\text{MLE}}$. Recall that the Bernoulli Distribution is discrete and has probability mass function: [3]

$$\mathbb{P}(X = k) = p^k(1-p)^{1-k}, \qquad k = 0, 1 \quad p \in [0, 1]$$

    (b) Assume that $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Compute $(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$ [4]

19. Having now practised finding MLE's, let us now rewrite equation 1 as [5]

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$
$$\implies y^{(i)} = \mathbf{x^{(i)}}^T \mathbf{w}$$

    And now the punchline: Let us assume that each of these data points have a degree of noise in them. Furthermore, we will assume that this noise is normally distributed with zero mean and variance σ^2:

$$y^{(i)} = \mathbf{x^{(i)}}^T \mathbf{w} + \epsilon^{(i)}, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

    Your task is write down the log-likelihood and maximum likelihood estimation objective and then solve for the MLE estimator $\hat{w}_{\text{MLE}}$. Note: you may assume the errors are independent and identically distributed.

§§§ Non-linearities

20. In this question we will explore the capacity of a linear model to fit **polynomial** relationships with a trick known as *Locally Weighted Regression*. [2]

    Locally Weighted Regression (LWR) is a non-parametric regression technique where weights are assigned to the data points based on their "closeness" to the query point. The weighted linear regression model minimizes a weighted version of the squared loss:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} w^{(i)} \left(y^{(i)} - \mathbf{x^{(i)}}^T \mathbf{w} \right)^2,$$

where the weights are given by:

$$w^{(i)} = \exp\left(-\frac{\|\mathbf{x}^{(i)} - \mathbf{x}\|^2}{2\tau^2}\right).$$

Here, $\tau > 0$ is a bandwidth parameter controlling the locality of the regression.

(a) Explain the difference between a parametric and non-parametric model. [1]

(b) Derive the closed-form solution for the weight vector $\mathbf{w}$ by solving the weighted least squares minimization problem. [2]

(c) For the dataset: [2]

$$\{(x^{(i)}, y^{(i)})\} = \{(-1, 1), (0, 0), (1, 1)\},$$

use Locally Weighted Regression with $\tau = 1$ to compute the predicted value $\hat{y}$ at $x = 0.5$

21. What is an n-dimensional line called? [1]

22. We will now see how it is possible to construct a linear decision boundary for data which looks as tangled as Figure Figure 2. [3]

The concentric circles have been randomly coloured, but the point of the task is to colour them correctly.

Your task is simple, submit a sketch or plot in 2 or 3 dimensions, that separates this data. This task may require some research into kernel methods, which project the data into higher dimensions and allow you to linearly separate the data in at least two of these dimensions

23. Finally, because all of the above methods are prone to overfitting, find the closed form solution of the multivariate least squares subject to L2 normalisation throttled by the hyperparameter lambda. [2]

$$w_{\text{ridge}} = \arg\max_{w} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

Note: Ridge regression is another term for L2 normalised Least Squares

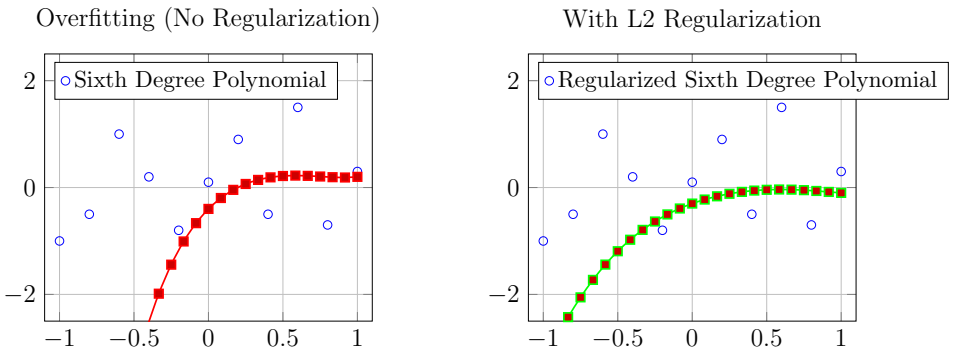

Figure 3: for conceptual benefit

§ Grade Table

Question:	1	2	3	4	5	6	7	8	9	10	11	12
Points:	4	3	3	2	4	2	4	6	3	3	2	5
Score:												

Question:	13	14	15	16	17	18	19	20	21	22	23	Total
Points:	2	5	5	1	2	7	5	7	1	3	2	81
Score:												

§ References

cs229 problem set 0; mathematics for machine learning (book); cs229 lecture notes; cs9417 lab code + tutorials. the diff eqns q were from a unsw math test. i created the matrix problems excluding q10. the remaining problems are simply classical.