UC Berkeley Self-Study/Review Notes: Machine Learning

Dun-Ming Huang

Contents

Chapter 1

DATA C100: Introduction to Modeling

1.1 Motivation

A **model** is an idealized representation of a system, which are mostly mathematical. Their mathematical properties lend us the computational opportunity to abstract a system in a computational space!
And in general, the machine learning works we perform in DATA C100 lend strength from constructions of models that allow us to predict new values from old data.
Outside the context of machine learning, there are a few categories of models that human history has used:

> - **Deterministic Physical (Mechanical) Models**, such as kinematics equations.
> - **Probabilistic Models**, which models how random processes can evolve.
> - **Statistical Models**, which associates variables via statistical analysis.
> - **Informal Models**, which are essentially stories or human-understandable descriptions of a complex phenomenon. Many pictographics might be an informal model.

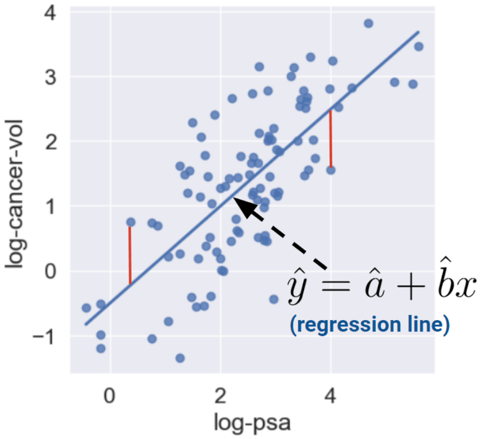
The introduction towards several categories of models only reinforce the idea that it is used for understanding the world as a complex phenomenon as well as providing predictions towards unseen cases.
Quite frequently, we would like to create models that are simple and interpretable, when we attempt to understand the association between different variables. But when we attempt to make extremely accurate preditcions, we would also risk providing an uninterpretable model whose complexity supports its performance.
Both directions of modeling in terms of interpretability have been used, and made their successes in each of their mission. Notably, complex models occur a lot in deep learning.

1.2 A Demonstration: Simple Linear Regression

1.2.1 What is a Regression Line?

Suppose we have a crowd of data points spread across a 2D plot, which we call a scatterplot, and we want to predict one dimension of the data point from another, then we may use a regression line:

Definition 1.2.1. Regression Line



A **Regression Line** is a linear model that attempts to predict a feature of a specific data-point via a linear combination of other features.

For now, let us simplify our description: "We would like to predict y by x".
The way we predict is by assuming there is a coefficient a and b such that:

$$y = a + bx$$

would accurately predict y for any provided x.
The most accurate parameter of this line would be denoted as $\hat{a}$ and $\hat{b}$, and the prediction following these best parameters would be denoted as $\hat{y}$, hence the regression line equation:

$$\hat{y} = \hat{a} + \hat{b}x$$

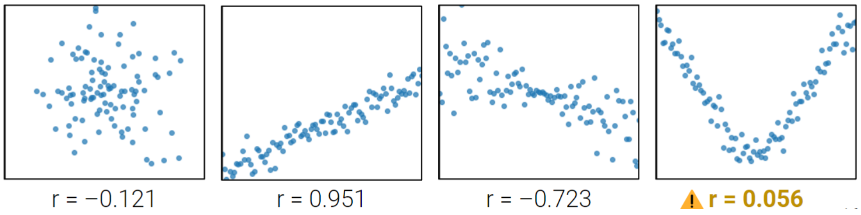
Each dataset that we attempt to apply a regression line onto has a specific statistic called **correlation**:

Definition 1.2.2. Pearson's Correlation Coeffcient

A **correlation** (denoted as r), formally known as the "Pearson's Correlation Coeffcient", quantifies how linearly associated two variables x and y may be via the following formula:

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

In other words, this is the average of the product of x and y, both measured in standard units.



At this point, you will probably have the following concerns:

- What does it mean when a parameter is "best" or "most accurate"?
- How do we obtain such parameters?

And here is where we deal with the mathematics of regression lines.

1.2.2 Model Selection: Simple Linear Regression

The **Simple Linear Regression** model is essentially a model using the regression line to predict the y of a datapoint for the x of a datapoint. In most definition, it's bound to only produce a two-dimensional regression line.
Such a model's prediction relies entirely on the value of its parameters. We recognize this by saying that Simple Linear Regression is a **parametric model**, described by its parameters a and b.
You might have noticed from the few graphs above that, the regression line doesn't accurately predict every single point. It is born from a set of points that don't necessarily form a line. Therefore, the best parameters that we provide it (again, called $\hat{a}$ and $\hat{b}$), are what we call the sample-based estimate of the inexistent, completely correct regression line.
Following that same logic, we noted the result of prediction as $\hat{y}$: our best estimate of the actual y of a new, unseen datapoint.

1.2.3 Quantifying Errors: Loss Functions

To quantify how "good" a prediction is, we would like to use a loss function:

Definition 1.2.3. Loss Function

A **loss function** is a function that characterizes the cost in predictions for choosing a set of parameters.
It quantifies how bad a prediction is for a single observation. The closer our prediction is to the actual observed value, the better the prediction, therefore, the lower the loss. Vice versa.

The choice of loss function affcets how we customize our model to adapt to mistakes. While it affects the accuracy of estimations, it would also decide the computational cost of estimation, as some loss functions are very costly to calculate for computers.
For Simple Linear Regression, we usually bother with two choices of loss functions:

Definition 1.2.4. Loss Functions of SLR

Squared Loss (L2 Loss)

$$L(y, \hat{y}) = (y - \hat{y})^2$$

This is a reasonable choice because there would be no loss when prediction is equal to observed value, and provide a lot of loss for predictions that are far from the observed values.

Absolute Loss (L1 Loss)

$$L(y, \hat{y}) = |y - \hat{y}|$$

This is a reasonable choice because there would be no loss when prediction is equal to observed value, and provide a fair amount of loss at a uniform sign (positive) for predictions that are far from the observed values (benefited from the absolute value).

Since we concern how costly our model's predictions are for the entire data set, the natural measure of how good a model is would be the average loss of model across all data points. This is also known as **empirical risk**:

Definition 1.2.5. Empirical Risk

The empirical risk is the average loss of a model across all data points. Minimizing empirical risk provides the optimal estimated parameters of a model for the corresponding loss function.
Mathematically expressed,

$$\hat{R}(\theta) = \frac{1}{n}\sum_{i} L(y_i, \hat{y}_i)$$

Here, θ represents the parameters of the model.

Let us inspect the case where we attempt to minimize the empirical risk with our loss being L2 loss function. In this case, we call our empirical risk **Mean Squared Error**.

Derivation 1.2.1: Estimated Parameters of Simple Linear Regression under MSE

To recall:

$$MSE(a,b) = \frac{1}{n}\sum_i (y_i - a - bx_i)^2$$

To minimize this function, we will find the conditions where the partial derivative of MSE with respect to every parameter is 0.
Minimizing a:

$$\begin{aligned}
\frac{\partial MSE}{\partial a} &= \frac{1}{n} \times \sum_i 2\frac{\delta}{\delta a}(y_i - a - bx_i)(y_i - a - bx_i) \\
&= \frac{1}{n} \times \sum_i -2(y_i - a - bx_i) \\
&= -\frac{2}{n}\sum_i (y_i - a - bx_i) \\
\frac{1}{n}\sum_i (y_i - \hat{y}_i) &= 0 \\
\sum_i (y_i - a - bx_i) &= \sum_i (y_i) - a - b\sum_i (x_i) \\
&= \overline{y} - a - b\overline{x} = 0 \\
a &= \overline{y} - b\overline{x}
\end{aligned}$$

Minimizing b:

$$\begin{aligned}
\frac{\partial MSE}{\partial b} &= \frac{1}{n} \times \sum_i 2\frac{\delta}{\delta b}(y_i - a - bx_i)(y_i - a - bx_i) \\
&= \frac{1}{n} \times \sum_i -2x_i(y_i - a - bx_i) \\
&= -\frac{2}{n}\sum_i x_i(y_i - a - bx_i) \\
\frac{1}{n}\sum_i x_i(y_i - \hat{y}_i) &= 0 \\
\frac{1}{n}\sum_i x_i(y_i - \hat{y}_i) - \frac{1}{n}\sum_i \overline{x}(y_i - \hat{y}_i) &= \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \hat{y}_i) \\
&= \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y} - b(x_i - \overline{x})) = 0 \\
\frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y}) &= b\frac{1}{n}\sum_i (x_i - \overline{x})(x_i - \overline{x}) \\
r_{x,y}\sigma_x\sigma_y &= b\sigma_x^2 \\
b &= r\frac{\sigma_y}{\sigma_x}
\end{aligned}$$

Conclusion:

$$\begin{cases} \hat{a} = \overline{y} - \hat{b}\overline{x} \\ \hat{b} = r\frac{\sigma_y}{\sigma_x} \end{cases}$$

So at last, we also find that the optimizing condition of SLR model would be:

$$\begin{cases} \frac{1}{n}\sum_i (y_i - \hat{y}_i) & = 0 \\ \frac{1}{n}\sum_i x_i (y_i - \hat{y}_i) & = 0 \end{cases}$$

Which can be interpreted respectively as that:

> - The residuals average to 0.
>
> - The residuals are orthogonal to the predictor variable.

Chapter 2

DATA C100: Constant Model

2.1 Defining the Constant Model

Now that we have finished reading about simple linear regression from the previous lecture, let us discuss a slightly simpler model.

Definition 2.1.1. Constant Model

Constant Model, also known as a summary statistic, summarizes the sample data by always predicting the same number for any data point.
Mathematically expressed,

$$\hat{y} = \theta$$

In other words, the estimated y is always a constant parameter θ, whatever the input might be.

2.2 Experimenting Different Loss Functions

A model may have several options for its loss functions, which comes with different advantages and disadvantages. For this constant model, let us explore the L2 and L1 loss functions, see how each brings a different condition of optimization and robustness.

2.2.1 Exploring L2 Loss: MSE

In the L2 case, let us recall that our empirical risk is defined as follows:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Therefore, for our constant model:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

Let us proceed onto its optimization by attempting to find the critical point of empirical risk:

Derivation 2.2.1: Optimization of L2 Empirical Risk for Constant Model

$$\frac{\partial R}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} 2\frac{\delta}{\delta\theta}(y_i - \theta)(y_i - \theta)$$
$$= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta)$$

And now, to optimize R,

$$\sum_{i=1}^{n}(y_i - \theta) = 0$$
$$\bar{y} - \theta = 0$$
$$\theta = \bar{y}$$

Therefore,

$$\hat{\theta} = \bar{y}$$

Let us enjoy some observations here.
First of all, the minimum MSE is thus:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sigma_y^2$$

, the variance of y.
Second of all, the estimated value of parameter θ is thus the mean of y. This means, provided an extreme outlier, the model will misbehave due to the mean being heavily influenced by some extreme outlier(s). Therefore, L2 Loss is not very robust (adaptative) towards the appearance of outliers.

2.2.2 Exploring L1 Loss: MAE

The L1 Loss Empirical Risk is also known as **Mean Absolute Difference**, which followed a similar naming logic to MSE.
Mathematically expressed,

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \theta)|$$

The optimization of such an empirical risk becomes interesting (matheamtically), due to the appearance of absolute value.
However, we may always characterize the absolute value as a piecewise function:

$$f(x) = |x - \theta| \to f(x) = \begin{cases} x - \theta, & x > \theta \\ 0, & x = \theta \\ \theta - x, & x < \theta \end{cases}$$

Let us exploit this in the following toil:

Derivation 2.2.2: Optimization of L2 Empirical Risk for Constant Model

$$|y_i - \theta| = \begin{cases} y_i - \theta, & y_i > \theta \\ 0, & y_i = \theta \\ \theta - y_i, & y_i < \theta \end{cases}$$

$$\frac{\delta}{\delta\theta}|y_i - \theta| = \begin{cases} 1, & y_i > \theta \\ 0, & y_i = \theta \\ -1, & y_i < \theta \end{cases}$$

$$\frac{\partial R}{\partial \theta} = \frac{1}{n}\Big(\sum_{\theta<y_i}(-1) + \sum_{\theta>y_i}(1)\Big)$$

$$\sum_{\theta<y_i}(-1) + \sum_{\theta>y_i}(1) = 0$$

$$\sum_{\theta>y_i}(1) = \sum_{\theta<y_i}(1)$$

Therefore, $\hat{\theta}$ must be the median of y.

The estimated value of parameter θ is thus the mean of y. This means, provided an extreme outlier, the model will not misbehave due to the median not easily influenced by some extreme outlier(s). Therefore, L1 Loss is more robust (adaptative) towards the appearance of outliers.

2.2.3 Summary of Loss Function Optimization

In summary, the process of finding optimization conditions (which we also call **estimating equation**) would follow:

1. Differentiate the empirical risk with respect to parameters.
2. Attempt to find the critical point of empirical risk for per parameter.
3. Perform the derivative test (which requires multivariable calculus in most occassions, and is not performed for the span of DATA C100 for that reason) to confirm that the critical point is a minima.

The multivariable perspective offers a much computationally heavy test for confirming whether a critical point is a minima, which would involve calculating plural higher order partial derivatives. For those who are interested, this is in-scope for MATH 53.

2.3 Transformation and Model Linearity

In some cases, we face how the predictor variable x is not linearly correlated with y, but a transformation of x might be.

For example, the trajectory of a baseball is mostly quadratic. If I'd like to predict its motion, it is best that I present a model whose shape is not linear, but rather, quadratic:

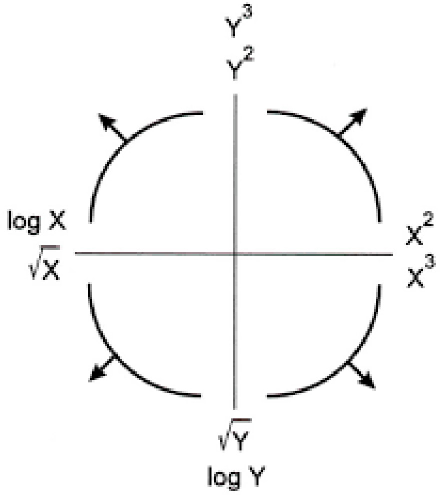
$$\hat{y} = \hat{a} + \hat{b}x^2$$

This happens frequently across datasets, but the question is: is the model still linear in this case?

The answer is yes. Linear Regression requires a regression line, decided as a linear combination of parameters. In this case, the predictor variable is being a non-linear term, but the regression equation is still linear with respect to each of the parameters.

As previously mentioned, models sometimes require transformations to behave better, fitting to the behaviour of the dataset more closely.
To decide what transformations seem optimal, we can employ the following figure:

Figure 2.3.1. Turkey-Mosteller Bulging Diagram



Each of these transformations suggest how to transform *x* and *y* via suggesting that, for the direction that data's shape currently bulges towards, we should transform *x* and/or *y* in that direction.
In this case, we are usually then given the choice to either transform *x* or *y*, or as priorly mentioned, perhaps both.

Chapter 3

DATA C100: Ordinary Least Squares

3.1 Multiple Linear Regression Model

We have seen Simple Linear Regression Model, which uses one parameter for a constant term and another parameter to introduce the predictor variable.
The model currently has one predictor variable.
What if we can use more? What if we need to use more? What if our model would benefit greatly because what we attempt to predict in nature requires two or more variables for a good prediction?
If so, such a model would have some regression equation whose shape is to a huge degree similarly shaped as below:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Let us attempt to vectorize the above equation. Suppose, we rewrite the above equation as follows:

$$\hat{y}^{(i)} = \begin{bmatrix} 1 & x_1^{(i)} & \cdots & x_p^{(i)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_p \end{bmatrix}$$

Then, we have successfully written the above equation in a matrix-vector multiplication form.
The anatomy of each term in the above row vector, $x_k^{(i)}$, is as follows:

> - $x^{(i)}$ stands for the i^{th} data point inside the dataset.
> - x_k stands for the k^{th} feature inside the dataset.
> - Therefore, $x_k^{(i)}$ is the k^{th} feature of i^{th} data point inside the dataset.
> - Among all columns, we appended another column of ones on the left of row vector to account for the need of intercept in regression line.

To vectorize this operation across numerous datapoints in the dataset, we may then formulate this model in terms of matrix-vector multiplication as follows:

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_k \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_p^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(k)} & \cdots & x_p^{(k)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_p \end{bmatrix}$$

And each term can then be respectively abbreviated into what is shown below:

$$\mathbb{Y} = \mathbb{X}\theta$$

Where we specifically name the matrix $\mathbb{X}$ containing the datapoints as the **design matrix**.

3.2 Optimization of Model: Least Squares Algorithm

3.2.1 Loss Function

The Loss function most frequently applied for such a model has to deal with a linear algebraic property called L2 Norm:

Definition 3.2.1. L2 Norm

The L2 Norm of a vector $\vec{x} \in \mathbb{R}^n$ is mathematically expressed as:

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

which occurs to be the magnitude of such vector $\vec{x}$.
We thus notate the distance of vectors $\vec{a}, \vec{b} \in \mathbb{R}^n$ as:

$$||a - b||_2$$

Our L2 loss function would be the distance between estimation and observed values, squared:

$$||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2$$

Therefore yielding the empirical risk:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

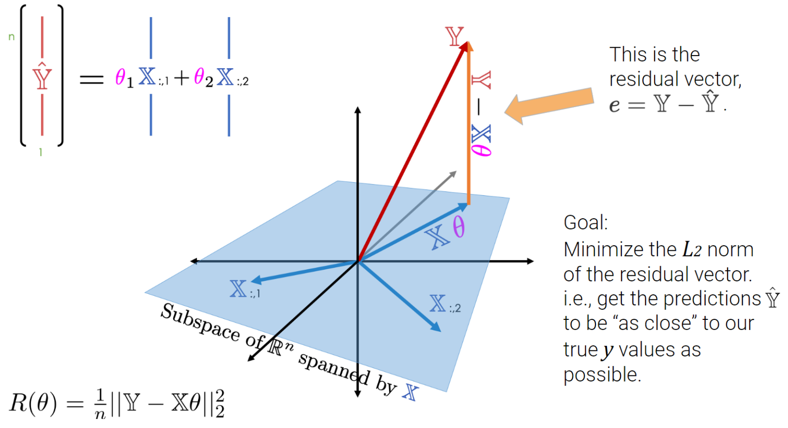
3.2.2 Optimization via Geometric Interpretation

We should first observe with our prior knowledge from EECS 16A (or alternatively MATH 54, just any college linear algebra introductory course), that since $\hat{Y}$ is a linear combination of the columns of $\mathbb{X}$ (as noted $\hat{Y} = \mathbb{X}\theta$),

$$\hat{\mathbb{Y}} \in span(\mathbb{X}) \subseteq \mathbb{R}^k$$

However, it is not necessary that $\mathbb{Y}$ belongs to the span of $\mathbb{X}$. Therefore, we see more clearly that our task is to minimize the distance between $\mathbb{Y} \notin span(\mathbb{X})$ and $\hat{\mathbb{Y}} \in span(\mathbb{X})$.
Let us observe a visualization from the DATA C100 Lecture Slides (since mine are underqualified and old):

Figure 3.2.1. Geometric Image of Residual in Multiple Linear Regression


n
$\hat{\mathbb{Y}}$
1
$= \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$
$\mathbb{Y}$
$\mathbb{Y} - \mathbb{X}\theta$
$\mathbb{X}\theta$
$\mathbb{X}_{:,1}$
$\mathbb{X}_{:,2}$
Subspace of $\mathbb{R}^n$ spanned by $\mathbb{X}$

This is the residual vector,
$e = \mathbb{Y} - \hat{\mathbb{Y}}$.

Goal:
Minimize the L_2 norm
of the residual vector.
i.e., get the predictions $\hat{\mathbb{Y}}$
to be "as close" to our
true y values as
possible.

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

This residual vector is essentially the shortest possible (minimized) when it is orthogonal to $\mathbb{X}\theta$ (which, in a 2D view, has to do with a property of right triangles called the Pythagorean Theorem).
There is a better intuition than Pythagorean Theorem, which is that the vector in $span(\mathbb{X})$ closest to $\mathbb{Y}$ must be its projection onto $span(\mathbb{X})$.
Either way, we will be introduced to a simplified situation:

To minimize

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

We require that

$$\mathbb{X}^T(\mathbb{Y} - \mathbb{X}\hat{\theta}) = 0$$

Such that the residual is orthogonal to $span(\mathbb{X})$.

Let us review its solution from EECS 16AB (or MATH 54, MATH 110):

Derivation 3.2.1: Least Squares Algorithm

$$\begin{aligned}\mathbb{X}^T(\mathbb{Y} - \mathbb{X}\hat{\theta}) &= 0 \\ \mathbb{X}^T\mathbb{Y} &= \mathbb{X}^T\mathbb{X}\hat{\theta} \\ \hat{\theta} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}\end{aligned}$$

The equation above that shows the optimal parameters is known as the **Normal Equation**.
This normal equation would only be useful when $\mathbb{X}^T\mathbb{X}$ is invertible. Determining whether $\mathbb{X}^T\mathbb{X}$ is invertible is not as difficult as it sounds, since $N(A^TA) = N(A)$. I will not showcase this proof here.
Beyond the scope of DATA C100, we should also see some remedy for such situations when attempting to solve this exact optimization problem and working with a non-invertible design matrix. It is also not in the scope of this note yet.

3.3 Performance Factors

Just like in previous models, the residuals should be uncorrelated with the predicted values $\hat{y}$.
To determine the correlation of variables in a multivariable perspective like in Multiple Linear Regression, we work with a new coefficient:

Definition 3.3.1. Coefficient of Determination

The coefficient of determination characterizes the correlation of variables in a Multiple Linear Regression model:

$$R^2 = \frac{\text{Variance of predicted values}}{\text{Variance of observed values}} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

Just like the Pearson's Correlation Coefficient (r in Simple Linear Regression), R^2 spans between 0 and 1 (except it is the absolute value of r that spans between 0 and 1, not r itself).

Chapter 4

DATA C100: Gradient Descent

4.1 Computational Minimization of Loss Function

Lorem Ipsum.

4.2 Gradient Descent Algorithm

Lorem Ipsum.

4.2.1 Interpretation of Gradients

Lorem Ipsum.

4.2.2 Gradient Descent Mathematically

Lorem Ipsum.

4.2.3 Stochastic Gradient Descent

Lorem Ipsum.

4.2.4 Convexity

Lorem Ipsum.

Chapter 5

DATA C100: Feature Engineering

5.1 Motivation

Lorem Ipsum.

5.2 One Hot Encoding

Lorem Ipsum.