COMP9417 Notes

Alperen Onur

May 2025

1 Regression

1.1 Supervised & Unsupervised Learning

The two types of machine learning algorithms fall under supervised or unsupervised learning. Supervised learning categorises algorithms with given labels and unsupervised algorithms categorise algorithms with no given labels.

1.2 Linear Regression

Regression predicts numerical values whereas **Classification** predicts discrete values. **Linear Regression** is a particular type of regression which models the relationship between an input variable and an output variable using a straight line,

$$\hat{y} = bx + c$$

where, $\hat{y}$ are our predicted labels based on our input features x. The goal in our case is to estimate the slope b and intercept c from the data. We make the following assumptions for linear regression,

- Linearity: The relationship between x and the mean of y is linear.
- Homoscedesticity: The variance of residual is the same for any value of x.
- Independence: Observations are independent of each other.
- Normality of residuals: For any fixed value of x, y is normally distributed.

1.3 Linear Regression Forumlation

In linear models, the outcome is a linear combination of attributes,

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = h(x)$$

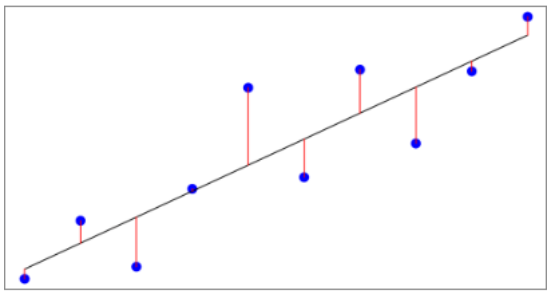
where θ_i is the weight from the observed training data for attribute i. The predicted value of the first training instance x_i is,

$$\hat{y}_1 = \theta_0 + \theta_1 x_{11} + \theta_2 x_{12} + \cdots + \theta_n x_{1n} = \sum_{i=0}^{n} \theta_i x_{1i} = x_1^T \theta = h(x_1).$$

Typically, x_0 is set to 1 and we define $x = [x_0, x_1, ..., x_n]^T$.

1.4 Minimizing Error

We can fit an infinite amount of lines to a dataset depending on what we define as the best fit criteria. The most popoular estimation model is "Least Square", also known as "Ordinary Least Squares" regression. This model attempts the minimize the difference between the predicted and actual values; minimizing the error.



The goal is the minimize the error over all input samples. The total error is defined as the sum of squared errors and searching for $n + 1$ parameters to minimize that. The sum of squared error for m samples is denoted as,

$$J(\theta) = \sum_{j=1}^{m} (y_j - \sum_{i=0}^{n} \theta_i x_{ji})^2 = \sum_{j=1}^{m} (y_j - x_j^T \theta)^2 = (\mathbb{y} - X\theta)^T (\mathbb{y} - X\theta)$$

where $J(\theta)$ is the loss function and $X, \mathbb{y}$ are

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ & & \vdots & & \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \qquad \mathbb{y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_m \end{bmatrix}$$

The loss function can also be generalised as averaging the squared error and minimizing it which results in the same θ,

$$J(\theta) = \frac{1}{m}\sum_{j=1}^{m}(y_j - x_j^T)^2.$$

This is typically referred to as the **mean squared error (MSE)**.

1.5 Least Squares Regression

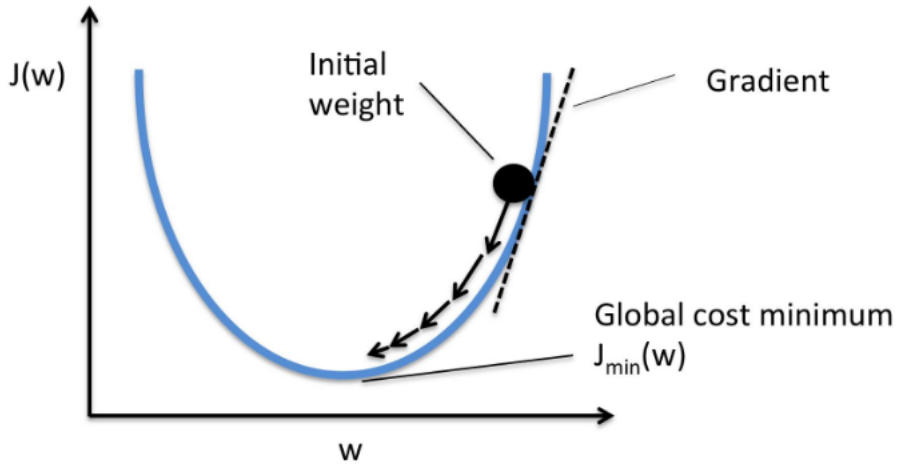
We need a method to estimate the parameters of the lost function as to minimize its overall cost. Less error means more accurate predictions. Computing $J(\theta)$ for different values of θ results in a convex function which has a single global minima. An algorithm to converge to this minima is **Gradient Decent**.

1.6 Gradient Decent

Gradient decent starts with some initial θ, and repeatedly performs an update,

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \alpha\frac{\partial}{\partial\theta_i}J(\theta_i^{(t)})$$

where α is the learning rate. In each iteration of the algorithm, it takes a "step" the size of α in the direction with the steepest increase in $J(\theta)$. To implement this algorithm, we'll need the partial derivative of the MSE.



For one sample of m samples, the cost function is,

$$J(\theta) = (y_j - h_\theta(x_j))^2 = (y_j - \sum_{i=1}^{n} x_{ji}\theta_i)^2 \tag{1}$$

$$h_\theta(x_j) = \sum_{i=1}^{n} x_{ji}\theta_i = x_j^T\theta \tag{2}$$

Now taking the derivative we get,

$$\frac{\partial}{\partial\theta_i}J(\theta) = -2(y_j - h_\theta(x_j))x_{ji}. \tag{3}$$

So for a single training sample, the update rule is,

$$\theta_i^{(t+1)} = \theta_i^{(t)} + 2\alpha(y_j - h_\theta(x_j))x_{ji}. \tag{4}$$

This update rule for squared distance is called **Least Mean Squares (LMS)**. For multiple samples, there are a couple methods used for updating the LMS rule: **Batch Gradient Decent** and **Stochastic Gradient Decent**.

1.7 Batch Gradient Decent

$$\theta_i^{(t+1)} = \theta_i^{(t)} + \alpha\frac{2}{m}\sum_{j=1}^{m}(y_j - h_{\theta_i^{(t)}}(x_j))x_{ji}.$$

For every i, replace the gradient with the sum of gradient for all samples until convergence (when θ is stabilized).

1.8 Stochastic Gradient Decent

For $j = 1$ to m:
[1] $\theta_i^{(t+1)} := \theta_i^{(t)} + 2\alpha(y_j - h_\theta(x_j))x_{ji}$ (for every i)
Repeat until algorithm converges. θ gets updated at each sample separately. This is much less costly than batch gradient decent but it may also never converge to a minimum.

1.9 Minimizing Square Error with Normal Equations

We can also find the minimum of $J(\theta)$ by explicitly taking its derivatives and setting them to zero. This is the closed form solution,

$$\frac{\partial}{\partial\theta}J(\theta) = 0 \tag{5}$$

$$J(\theta) = (\mathbb{y} - X\theta)^T(\mathbb{y} - X\theta) \tag{6}$$

$$\frac{\partial}{\partial\theta}J(\theta) = -2X^T(\mathbb{y} - X\theta) = 0 \tag{7}$$

$$X^T(\mathbb{y} - X\theta) = 0 \tag{8}$$

$$\theta = (X^TX)^{-1}X^T\mathbb{y} \tag{9}$$

1.10 Minimizing Square Error with Probabilistic Interpretation

We can write the relationship between an input variable x and output variable y as,

$$y_j = x_j^T\theta + \epsilon_j$$

where ϵ_j is an error term (random noise). If we assume all ϵ_j are independent and identically distributed according to the Gaussian distribution $\epsilon_j \sim N(0, \sigma^2)$ then,

$$p(\epsilon_j) = \tfrac{1}{\sqrt{2\pi\sigma^2}}\exp(-\tfrac{\epsilon_j^2}{2\sigma^2}).$$

This implies that, $P(\epsilon_j) = P(y_j|x_j\,;\theta) = \tfrac{1}{\sqrt{2\pi\sigma^2}}\exp(-\tfrac{(y_j - x_j^T\theta)^2}{2\sigma^2})$. So we want to estimate θ such that we maximize the probability of output y given input x over all m training samples. Mathematically this is,

$$\mathcal{L}(\theta) = P(\mathbb{y}|X\,;\theta)$$

where $\mathcal{L}(\theta)$ is the likelihood function. Since we assumed independence over ϵ_j, then each of our training samples are independent from each other. Then it follows that,

$$\mathcal{L}(\theta) = \prod_{j=1}^{m} P(y_j|x_j\,;\theta) \tag{10}$$

$$= \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}}\exp - \frac{(y_j - x_j^T\theta)^2}{2\sigma^2}. \tag{11}$$

This is called the **maximum likelihood**. If we want to find a θ that maximizes $\mathcal{L}(\theta)$, we can also maximize any strict increasing function of $\mathcal{L}(\theta)$. If we choose to maximize the **log likelihood** $\ell(\theta)$,

$$\ell(\theta) = \log \mathcal{L}(\theta) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y_j - x_j^T\theta)^2}{2\sigma^2} \tag{12}$$

$$= m \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{\sigma^2} \frac{1}{2} \sum_{j=1}^{m} (y_j - x_j^T\theta)^2. \tag{13}$$

So maximizing $\ell(\theta)$ is equal to minimizing $\sum_{j=1}^{m}(y_j - x_j^T\theta)^2$ which is the MSE.

2 Statistical Techniques for Data Analysis

2.1 Sampling

Sampling is a way to draw conclusions about a population without having to measure the entire population. For groups that are fairly homogeneous, we do not need to collect alot of data. For populations with alot of irregularities, we need to either take measurements from the entire group or find a some other way to get a good idea of the groups trends without having to do so. Sampling gives us a way to do this!

We want from our sampling method to have as little bias that we can account for and if the chance of obtaining an unrepresentative sample is high then we can choose to not draw conclusions. The chance of an unrepresentative sample decreases with the size of the sample.

2.2 Estimation

Estimation refers to the process by which one makes inferences about a population based on information obtained from a sample. A "good" estimate means that the estimator is correct on average. If the expected value of the estimator equals the true population parameter then it is said to be an unbiased estimator. The following are a few examples of different estimators,

- Sample mean: $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Sample median: The middle value when the sample data is ordered. Less affected by outliers; biased.
- Sample Variance $s^2 = \frac{1}{N-1} \sum_i (x_i - m)^2$
- Sample Standard Deviation: $s = \sqrt{s^2}$ Indicates the average distance of each sample from the mean; biased.

- Sample Range: Provides a simple measure of data spread by taking the difference between the maximum and minimum values; biased.

2.3 Covariance

Covariance is the measure of the relationship between two random variables,

$$cov(x, y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{N-1} = \frac{(\sum_i x_i y_i) - N\overline{xy}}{N-1}$$

2.4 Correlation

Correlation is a measure to show how strongly a pair of random variables are. The Pearson correlation between x and y is,

$$r = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}.$$

This only captures linear relationships between two variables. The Pearson correlation can range between -1 and 1. A value close to 1 shows high values of x are associated with high values of y and low values of x are associated with low values of y. Generally this means that the scatter is low. A value near 0 indicated there is no association; large scatter. A value close to -1 suggests a strong inverse association between x and y. Correlation is a quick way of checking whether there is some linear association between two variables x and y where the sign tells you the direction of that association.

3 Univariate Linear Regression

In order to find the parameters we take the partial derivatives, set them to 0 and solve for θ_0. This will lead to,

$$\theta_1 = \frac{cov(h, w)}{var(h)} \tag{14}$$

$$\theta_0 = \overline{w} - \theta_1 \overline{h}. \tag{15}$$

3.1 Linear Regression Intuitions

Adding a constant to all x values affects only the intercept but not the regression coefficient. We can then zero-center the x values by subtracting the mean of x in which case the intercept is equal to the mean of y. Similarly we can subtract the mean of y

from all y values to achieve a zero intercept, without changing the regression problem in an essential way. Another important point is that the sum of the residuals of the MSE makes linear regression susceptible to outliers far from the regression line.

3.2 Multiple Regression

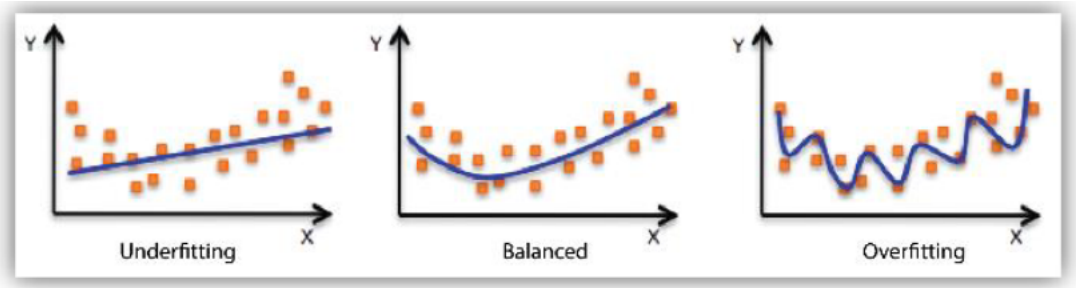
Often we need need to model the relationship of y to several other variables. Similar to univariate regression we can model this as,

$$\hat{w} = \theta_0 + \theta_1 h + \theta_2 b$$

and minimize the sum of the square residuals like so,

$$\sum_{j=1}^{m} (w_j - (\theta_0 + \theta_1 h_j + \theta_2 b_j))^2.$$

These types of regression models tend to produce curves rather than lines. So we can predict an output by adding different amounts of sample terms with each term increasing the overall degree of the models polynomial factor.



3.3 Regularisation

To control for overfitting or underfitting we introduce the concept of **regularisation**. Regularisation applies addition constraints the the weight vectors of a model. The regularised form for linear regression could be,

$$J(\theta) = \sum_{j=1}^{m} (y_j - h_\theta(x_j))^2 + \lambda \sum_{i=1}^{n} \theta_i^2.$$

The multiple least square regression problem is an optimisation problem and can be written as,

$$\theta^* = \arg\min \theta \ (y - X\theta)^T (y - X\theta)$$

with the regularised version of this being,

$$\theta^* = \arg\min \theta \ (y - X\theta)^T(y - X\theta) + \lambda||\theta||^2$$

where $||\theta||^2 = \sum_i \theta_i^2$ is the square norm of the vector θ; the dot product $\theta^T\theta$.

3.4 Ridge Regression

The regularised problem has a closed-form solution,

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

where I denotes the identity matrix. This adds λ amount of regularisation to the diagonal of $X^T X$. This is known as Ride Regression.

3.5 LASSO Regression

LASSO regression replaces the ride regression term $\sum_i \theta^2$ with the sum of absolute weights $\sum_i |\theta|$. LASSO regression favours more sparse solutions.

4 Train, Validation & Test Data

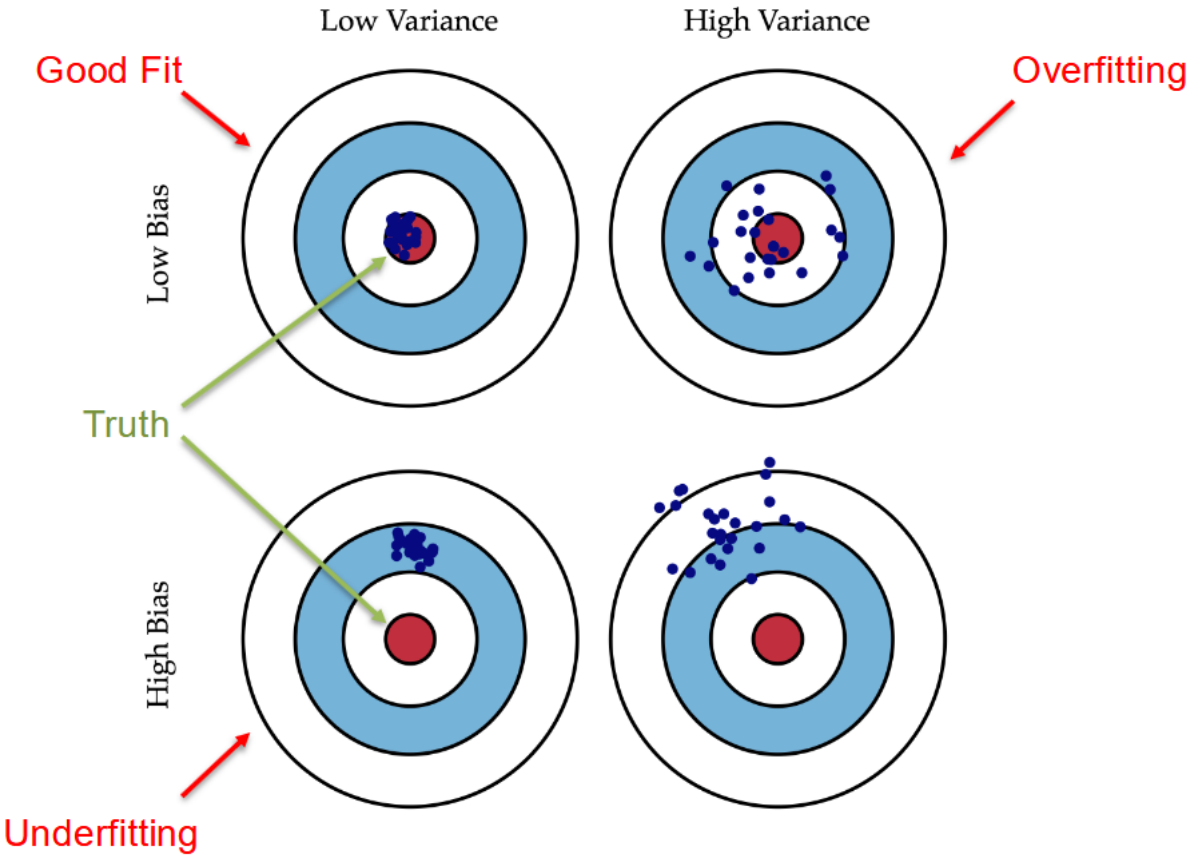
Train data is the data we use to learn our model and its parameters. Validation data is the unseen data by our model used to provide an unbiased evaluation of a model fit on the training dataset while tuning the models hyper-parameters. The test data is the data we use to test the model and shows how well our model generalises.

4.1 Model Selection

There are 3 ways to reduce complexity of a model,

- Subset-selection: search over a subset lattice such that each subset results in a new model and select one of those models.
- Shrinkage: Use regularisation to set coefficient of models to 0.
- Dimensionality reduction: Project points to a lower dimensional space.

4.2 Bias-Variance Tradeoff



Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Models with high bias pay very little attention to the training data and oversimplifies the model. **Variance** is the variability of the model for a given data point or a value which tells us the spread of data. Models with high variance pays a lot of attention to the training data but does not generalise on the data which has not been seen before. Underfitting means high bias and low variance. Overfitting means our model has captured a lot of noise along the underlying pattern in the data.

4.3 Bias Variance Decomposition

Whenw e assume $y = f + \epsilon$ and we estimate f with $\hat{f}$, then the expectation of error is,

$$E[(y - f)^2] = (f - E[\hat{f}])^2 + \text{Var}(\hat{f}) + \text{Var}(\epsilon)$$

and so the MSE can be written as,

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \text{irreducible error}.$$

Irreducible error is the inherent uncertainty associated with a natural variability in a system. It can not be reduced since it is due to unknown factors. Reducible error is error that can and should be minimised further by adjustments to the model.

If our model is simple and has few parameters, then it may have high bias and low variance. On the other hand, if our model has a large number of parameters then it's going to have high variance and low bias. So we need to find a good balance without overfitting or underfitting to the data.