Probability for Deep Learning
Part 3

Johar M. Ashfaque

I. LINEAR REGRESSION

The term "regression" is due to Sir Francis Galton. Regression is a method for studying the relationship between a response variable Y and a covariate X. The covariate is also called a predictor variable or a feature. There can be one or more covariates. The data are of the form

$$(Y_1, X_1), ..., (Y_n, X_n).$$

One way to summarize the relationship between X and Y is through the regression function

$$r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)dy.$$

A. Simple Linear Regression

The simplest version of regression is when X_i is simple (a scalar not a vector) and $r(x)$ is assumed to be linear

$$r(x) = \beta_0 + \beta_1 x.$$

This model is called the simple linear regression model. Let $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$. Then

$$\mathbb{E}(\epsilon_i|X_i) = 0.$$

Let $\sigma^2(x) = \mathbb{V}(\epsilon_i|X = x)$. We will make the further simplifying assumption that $\sigma^2(x) = \sigma^2$ does not depend on x. We can thus write the linear regression model as follows.

Definition I.1 *(The Linear Regression Model)*

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

$$\mathbb{E}(\epsilon_i|X_i) = 0$$

and

$$\sigma^2(x) = \mathbb{V}(\epsilon_i|X_i).$$

The unknown parameters in the model are the intercept β_0 and the slope β_1 ad the variance σ^2 Let $\hat{\beta_0}$ and $\hat{\beta_1}$ denote the estimates of β_0 and β_1 respectively. The fitted line is defined to be

$$\hat{r}(x) = \hat{\beta_0} + \hat{\beta_1}(x).$$

The predicted values or fitted values are $\hat{Y}_i = \hat{r}(X_i)$ and the residuals are defined to be

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta_0} + \hat{\beta_1} X_i).$$

The residual sums of squares or RSS is defined by

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

The quantity RSS measures how well the fitted line fits the data.

Definition I.2 *The least squares estimates are the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize*

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Theorem I.3 *The least squares estimates are given by*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n.$$

An unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

II. GRAPHICAL MODELS

Graphical models are a class of multivariate statistical models that useful for representing independence relations. Graphical models often require fewer parameters and may lead to estimators with smaller risk. There are two main types of graphical models: undirected and directed.

An undirected graph $\mathcal{G} = (V, E)$ has a finite set V of vertices and a set E of edges consisting of a pair of edges. The vertices correspond to random variables X, Y, Z,... and edges are written as unordered pairs. For example, $(X, Y) \in E$ means that X and Y are joined by an edge. Two vertices are adjacent denoted $X \sim Y$ if there is an edge between them. A graph is complete if there is an edge between every pair of vertices. A subset $U \subset V$ of vertices together with their edges is called a subgraph. If A, B and C are three distinct subsets of V, we say that C separates A and B if every path from a variable in A to a variable in B intersects a variable in C.

Directed graphs are similar to undirected graphs except that there are arrows between vertices instead of edges. Like undirected graphs, directed graphs can be used to represent independence relations. A directed graph $\mathcal{G}$ consists of a set of vertices V and an edge set E of ordered pairs of variables. If $(X, Y) \in E$ then there is an arrow pointing from X to Y. If an arrow connects two variables X and Y in either direction, we say that X and Y are adjacent. If there is an arrow from X to Y then X is a parent of Y and Y is the child of X. The set of all parents of X is denoted $\pi(X)$. A directed path from X to Y is a set of vertices beginning with X and ending with Y such that each pair is connected by an arrow and all the arrows point in the same direction. A sequence of adjacent vertices starting with X and ending with Y but ignoring the direction of the arrows is called an undirected path. X is an ancestor of Y if there is a directed path from X to Y. We also say that Y is a descendent of X. A directed path that starts and ends at the same variable is called a cycle. A directed graph is acyclic if it has no cycles. In thie case we say that the graph is a directed acyclic graph (DAG).

III. LOG-LINEAR MODELS

Log-linear models are useful for modelling multivariate discrete data. There is a strong connection between log-linear models and undirected graphs.

A. The Log-Linear Model

Let $X = (X_1, ..., X_m)$ be a random vector with probability function

$$f(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, ..., X_m = x_m)$$

where $x = (x_1, ..., x_m)$. Let r_j be the number of values that X_j takes. Without loss of generality, we can assume that $X_j \in \{0, 1, ..., r_j - 1\}$. Suppose now that we have n such random vectors. We can think of the data as a sample from a multinomial with $N = r_1 \times r_2 \times ... \times r_m$ categories. The data can be represented as counts in a $r_1 \times r_2 \times ... \times r_m$ table. Let $p = (p_1, ..., p_N)$ denote the multinomial parameter. Let $S = \{1, ..., m\}$. Given a vector $x = (x_1, ..., x_m)$ and a subset $A \subset S$, let $x_A = (x_j : j \in A)$. For example, if $A = \{1, 3\}$ then $x_A = (x_1, x_3)$.

Theorem III.1 *The joint probability function $f(x)$ of a single random vector $X = (X_1, ..., X_m)$ can be written as*

$$\log f(x) = \sum_{A \subset S} \psi_A(x)$$

where the sum is over all subsets A of $S = \{1, ..., m\}$ and the ψ's satisfy the following conditions

- *$\psi_\emptyset(x)$ is a constant*
- *For every $A \subset S$, $\psi_A(x)$ is only a function of x_A and not the rest of the x_j's*
- *If $i \in A$ and $x_i = 0$ then $\psi_A(x) = 0$*

The formula

$$\log f(x) = \sum_{A \subset S} \psi_A(x)$$

is called the log-linear expansion of f. Note that this is the probability function for a single draw. Each $\psi_A(x)$ will depend on some unknown parameters β_A. Let $\beta = (\beta_A : A \subset S)$ be the set of all these parameters. We will write $f(x) = f(x; \beta)$ when we want to estimate the dependence on the unknown parameters β.

In terms of the multinomial, the parameter space is

$$\mathcal{P} = \left\{ p = (p_1, ..., p_N) : p_j \geq 0, \sum_{j=1}^{N} p_j = 1 \right\}.$$

This is the $N - 1$ dimensional space. In the log-linear representation, the parameter space is

$$\Theta = \left\{ \beta = (\beta_1, ..., \beta_N) : \beta = \beta(p), p \in \mathcal{P} \right\}$$

where $\beta(p)$ is the set of β values associated with p. The set Θ is a $N - 1$ dimensional surface in $\mathcal{R}^n$.