Elements of Statistical Learning Notes

Daniel Ruiz

May 2019

Contents

Notation:
We'll typically denote input variable by the symbol X. If X is a vector, its components can be accessed by subscripts X_j.

Quantitative outputs will be denoted by Y, and qualitative outputs by G.

We use uppercase letters X, Y or G when referring to the generic aspects of a variable. Observed values are written in lowercase; hence the ith observed value of X is written as x_i (x_i is again a scalar or vector).

Matrices are represented by bold uppercase letters. An example of this would be a set of N input p-vectors x_i, i=1,..,N (each x_i is a vector of p components) would be represented by $N \times p$ matrix $\mathbf{X}$.

We use a hat to signify a prediction and withhold a hat to signify a measurement. Hence, if Y denotes a measured value, $\hat{Y}$ would represent our predicted value.

1 CH.1: Primers on Background Material

2 CH.2: Overview of Supervised Learning

2.1 Introduction

Definition 2.1: Inputs

These are the set of variables that are measured / preset. In the statistical and pattern recognition literature, this also goes by the name of predictors *and* features*. Classically, they are called the* independent variables*.*

Definition 2.2: Output

The inputs influence one or more of these. These are the results. Similarly, output are also referred to as the responses *or classically as the* dependent variables*.*

Definition 2.3: Supervised Learning

Supervised learning *is the process of using inputs to predict the value of outputs.*

2.2 Variable Types and Terminology

Definition 2.4: Categorical / Discrete Variables

Outputs will generally vary in their nature. An example of this is the qualitative *character of outputs such as eye color assuming values in a finite set* $\mathcal{G} = \{blue, brown, green\}$*. Such a class isn't equipped with any explicit ordering and often descriptive labels rather than numbers are used to denote classes. Qualitative variables are also referred to as* categorical *or* discrete *variables as well as* factors*.*

Definition 2.5: Regression and Classification

Distinction in output type has led to a naming convention among prediction tasks. We use the term regression *to generally refer to predicting quantitative outputs whereas* classification *is reserved for qualitative outputs.*

Definition 2.6: Ordered Categorical

A third variable type that deals with notions such as small, medium *and* large*, with an ordering between the values, but no metric is appropriate (difference between medium and small need not be same as between large and medium).*

A cursory description of the learning task is: Given the value of an input vector X, make a good prediction of the output Y, denoted by $\hat{Y}$. Hence, if Y takes on values in $\mathbb{R}$, then so should $\hat{Y}$. Follows analogously for categorical outputs.

Definition 2.7: Training Data

To construct prediction rules, we require data and often a lot of it. Hence, we suppose that we have available a set of measurements (x_i, y_i) *or* (x_i, g_i)*, i = 1, .., N, which is known as the* training data*.*

2.3 Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors

Brief statement on *Least Squares* and *K-nearest Neighbours*:

Least Squares: Large assumption about the structure but stable. Possibly may give inaccurate predictions.

K-nearest: Mild structural assumptions with predictions often accurate but possibly unstable.

2.3.1 Linear Models and Least Squares

Definition 2.8: Linear Model and Bias

Given some vector of inputs $X^T = (X_1, X_2, .., X_p)$, we predict the output Y through the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \tag{2.1}$$

$\hat{\beta}_0$ is the intercept, which is also known as the bias in machine learning. We note that (2.1) is a general expression for a linear map, taking $X \mapsto \hat{Y}$. It's convenient to absorb the 1 into the X vector through a redefinition. Hence, defining the vector $\hat{\beta} = (\beta_0, \beta_1, .., \beta_p)$, we can alternatively express our linear model (2.1) as:

$$\hat{Y} = X^T \hat{\beta} \tag{2.2}$$

Definition 2.9: Hyperplane in Input-Output Space

In the (p+1)-dimensional input-output space S, (X, $\hat{Y}$) would represent a hyperplane. That is,

$$(X, \hat{Y}) = \{(x, y) \in S |\ x \in X, y = X^T \beta\} \tag{2.3}$$

Proposition 2.1: Gradient of the Linear Model

Suppose we view the function over p-dimensional input space by $f(X) = X^T \beta$. Then the gradient $f'(X)$ is given by:

$$f'(X) = \beta \tag{2.4}$$

We note that β is the vector in input space that points in the direction of steepest ascent. This fact is observed from considering the directional derivative. A standard exercise is showing that it is maximized in the direction of the gradient.

Definition 2.10: Residual Sum of Squares for Linear Model, RSS

Suppose that we have a set of N data points (i.e pairing of observed input-outputs (x, y)). This set would be given by $\{(x_i, y_i) | i \in \mathbb{Z}_N\}$. We define the residual sum of squares as follows:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{2.5}$$

We note that generically x_i and β are p-vectors, hence the notation. Notice that RSS essentially measures the level of deviation from our observed value y_i with what would be predicted by the linear model $x_i^T \beta$ for *some β. The goal therefore becomes in finding a suitable choice for β that minimizes RSS.*

Proposition 2.2: RSS Minimization

Given the residual sum of squares, $RSS(\beta)$ as defined Def 2.10 and provided that $\mathbf{X}^T\mathbf{X}$ is invertible, then RSS is minimized when β takes on the following value:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{2.6}$$

Where we have defined $\mathbf{X}$ as the $N \times p$ matrix with x_i forming the i^{th} row of the matrix. In essence, $\mathbf{X}_{ij} = (x_i)_j$. Hence, the fitted value at the i^{th} input x_i is given by $\hat{y}_i = \hat{y}(x_i) = x_i^T\beta$. Insofar as computation goes, we have the identity:

$$\mathbf{X}^T\mathbf{X} = \sum_k x_k x_k^T \tag{2.7}$$

2.3.2 Nearest-Neighbour Methods

This method aims to utilize the observations in the training set that are closest to some fixed point x in the input space. In essence, if we wanted to predict the output of some point in our input space, we search the immediate vicinity so as to establish whether the local configuration of observed data would favour a prediction at the desired point to go in a particular direction. Suppose that there were two outputs of {blue, red} and I selected some point completely surrounded by blue data. This method is constructed so as to predict the output to be blue as well.

Definition 2.11: K-Nearest Neighbours

Let (x_i, y_i) represent an input-output pair from our training sample (observed data). Then the k-nearest neighbours fit for $\hat{Y}$ is defined as follows:

$$\hat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i \tag{2.8}$$

Where $N_k(x)$ denotes the k closest points in the training sample lying in the neighbourhood of x. We note that invoking a notion of closeness means we have assumed a metric space. Typically, we will assume the metric to be the Euclidean distance. Note

2.4 Statistical Decision Theory

Definition 2.12: Loss Function

We aim to define a function f(X) for predicting Y, given values of the input X. Hence, we require the notion of a loss function $L(Y, f(X))$ for penalizing errors in prediction.

Definition 2.13: Squared Error Loss Function

The most common and convenient loss function is squared error loss:

$$L(Y, f(X)) = (Y - f(X))^2 \tag{2.9}$$

Proposition 2.3

Let $X \in \mathbb{R}^p$ denote a real valued random input vector (of dimension p), and $Y \in \mathbb{R}$ a real valued random output vector with joint distribution Pr(X,Y). The criterion for choosing f will now be demonstrated. We define the expected (squared) prediction error by

$$EPE(f) = E[(Y - f(X))^2] \tag{2.10}$$

where the expectation is with respect to the joint distribution. We will show that the solution to the function f that minimizes EPE is given by the regression function $f(x) = E[Y|x = x]$.

Proof. We first define the joint density $pr(x, y)$ to satisfy $Pr(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} pr(u, v)\ \mathrm{d}u\ \mathrm{d}v$. We also note from the definition of conditional density, we have that $pr(x, y) = pr(y|x)pr(x)$ where $pr(x)$ denotes the marginal density of the random variable X. We therefore have

$$\begin{aligned} EPE(f) &= \int\int [y - f(x)]^2 pr(x, y)\ \mathrm{d}x\ \mathrm{d}y && (2.11)\\ &= \int\bigg(\int [y - f(x)]^2 pr(y|x)\ \mathrm{d}y\bigg) pr(x) dx && (2.12) \end{aligned}$$

We observe that EPE is a functional of f and can therefore be subject to variational techniques. For convenience, we define $EPE(f) = \int \mathcal{L}_{EPE}[f]\ dx$ where $\mathcal{L}_{EPE}$ is explicitly defined by

$$\mathcal{L}_{EPE}(f) = pr(x)\int [y - f(x)]^2 pr(y|x)\ \mathrm{d}y. \tag{2.13}$$

Observe that EPE is bounded below since it's argument is positive definite and therefore extremizing should yield a minimization solution[a]. Observe that the Euler-Lagrange equation for this system is necessarily:

$$\frac{\delta}{\delta f} EPE = \frac{\partial \mathcal{L}_{EPE}}{\partial f} = 0 \tag{2.14}$$

We now compute the partial derivative of $\mathcal{L}_{EPE}$ with respect to f:

$$\begin{aligned} \frac{\partial \mathcal{L}_{EPE}}{\partial f} &= 2\ pr(x)\int [y - f(x)] pr(y|x)\ \mathrm{d}y && (2.15)\\ &= 2\ pr(x)\bigg(\int y\ pr(y|x)\ \mathrm{d}y - f(x)\int pr(y|x)\ \mathrm{d}y\bigg) && (2.16)\\ &= 2\ pr(x)\big(E[Y|X = x] - f(x)\big), && (2.17) \end{aligned}$$

where we have used the fact that $\int\ pr(y|x)\ \mathrm{d}y = 1$. Hence, we can therefore conclude that

$$f(x) = E[Y|X = x] \tag{2.18}$$

□

[a] Note that $pr(x)$ and $pr(y|x)$ are probability densities, which are always positive definite.

Proposition 2.4

An alternative procedure to Proposition 2.3 involves a pointwise minimization, but this will turn out to be an effectively equivalent proof. Let $X \in \mathbb{R}^p$ denote a real valued random input vector (of dimension p), and $Y \in \mathbb{R}$ a real valued random output vector with joint distribution Pr(X,Y). We define the expected (squared) *prediction error by*

$$\begin{aligned}EPE(f) &= E[(Y - f(X))^2] \\ &= \int \left(\int (y - f(x))^2 pr(y|x) dy \right) pr(x) dx \\ &= \int E_{Y|X}[(y - f(x))^2 | x] pr(x) dx \\ &= E_X\left[E_{Y|X}\left[(Y - f(X))^2 | X\right]\right].\end{aligned} \tag{2.19}$$

Then, it's sufficient to minimize this pointwise, which is essentially what was performed in Proposition 2.3. In essence, the solution to minimizing EPE is given by

$$f(x) = \underset{c \in \mathbb{R}}{\operatorname{argmin}}\ E_{Y|X}\left[(Y - c)^2 | X = x\right]. \tag{2.20}$$

One can observe this being true as $E_{Y|X}$ and $pr(x)$ are positive definite. The expectation, E_X thereby preserves this ordering. The solution is therefore given by

$$f(x) = E[Y|X = x] \tag{2.21}$$

Proof. We define $g(c, x) := E_{Y|X}\left[(Y - c)^2 | X = x\right]$. We observe that g is clearly continuous and differentiable in c. It's also positive definite and thereby has a minima. Simple differentiation gives us

$$\frac{\partial g}{\partial c} = 2 \int (y - c)\ pr(y|x)\ \mathrm{dy}. \tag{2.22}$$

Hence,

$$c_{min} = \int y\ pr(y|x)\ \mathrm{dy}, \tag{2.23}$$

which satisfies $0 = \frac{\partial g}{\partial c}(c_{min})$, thereby setting $f(x) = E[Y|X = x]$. □

Nearest Neighbours Implementation

Nearest neighbours tries to directly implement this recipe with training data. In essence, we essentially ask for the average of all y_i's with input $x_i = x$. One can settle for

$$\hat{f}(x) = Ave(y_i | x_i \in N_k(x)), \tag{2.24}$$

where Ave denotes average and $N_k(x)$ is a neighbourhood containing k points in $\mathcal{T}$ closest to x. We note that there are two approximation being used here:

1) The Expectation is approximated by an averaging over sample data.
2) Conditioning at a point is relaxed to conditioning on some region close to the target point.

Under mild regularity conditions of Pr(x,y), one can show that as $N, k \to \infty$ such that $k/N \to 0$, then $\hat{f}(x) \to E(Y|X = x)$.

Proposition 2.5

Let $X \in \mathbb{R}^p$ denote a real valued random input vector (of dimension p), and $Y \in \mathbb{R}$ a real valued random output vector with joint distribution Pr(X,Y). Suppose that we approximate f(x) to be a linear function, given below by

$$f(x) \approx x^T \beta \tag{2.25}$$

Then, EPE is given by $EPE(\beta) = E[(Y - X^T\beta)^2]$. We will show that the solution to β is given by $\beta = (E[XX^T])^{-1}E[XY]$.

Proof. Similarly, we express our expectation as before

$$EPE(\beta) = \int\int [y - x^T\beta] pr(x, y) \text{ d}x \text{ d}y \tag{2.26}$$

We aim to minimize as before, this time noting that EPE is regular function of β. Hence, we aim to solve for $\nabla(EPE) = 0$ (set the gradient to zero). We compute

$$\frac{\partial}{\partial \beta_i} EPE(\beta) = 2\int\int (y - x^T\beta)x_i \text{ d}x \text{ d}y \tag{2.27}$$

Hence, the gradient is given below by (Where we have assumed our space to be Euclidean)

$$\nabla(EPE) = 2\int\int (y - x^T\beta)x \text{ d}x \text{ d}y \tag{2.28}$$

We now press onwards to solve for β

$$\begin{aligned} 0 &= \int\int yx \text{ d}x \text{ d}y - \int\int xx^T\beta \text{ d}x \text{ d}y && (2.29)\\ &= E[XY] - \left(\int\int xx^T \text{ d}x \text{ d}y\right)\beta && (2.30)\\ &= E[XY] - E[XX^T]\beta && (2.31)\\ & && (2.32)\end{aligned}$$

We emphasize that $x^T\beta$ is a scalar and we can therefore move the x term across it. We have also moved β outside the integral since it has no x or y dependence. Note that $E[XX^T]$ is a matrix valued expectation (in particular, it is the cross-correlation matrix of $\mathbf{X}$ with itself) which we assume is invertible and therefore yields the unique solution given by

$$\beta = (E[XX^T])^{-1}E[XY] \tag{2.33}$$

□

Least Squares and K-nearest Neighbours

Both Least-Squares and k-nearest neighbours end up approximating conditional expectations by averages.

- *Least-Squares assumes that f(x) is well approximated by a globally linear function.*
- *k-nearest neighbours assumes that f(x) is well approximated by a locally constant function.*

Additive Models

Many of the techniques that we will encounter in this book are model based, although more flexible than the rigid linear model. For instance, additive models assume that

$$f(X) = \sum_{j=1}^{p} f_j(X_j) \tag{2.34}$$

This retains additivity of linear model, but each coordinate function f_j is arbitrary.

Proposition 2.6

Let $X \in \mathbb{R}^p$ denote a real valued random input vector (of dimension p), and $Y \in \mathbb{R}$ a real valued random output vector with joint distribution Pr(X,Y). We now instead consider the following loss function: $L_1 := |Y - f(X)|$. We define the expected absolute prediction error by

$$EPE(f) = E\big[|Y - f(X)|\big]. \tag{2.35}$$

The minimizing solution is given by

$$\hat{f}(x) = median(Y|X = x) \tag{2.36}$$

Proof. We expand the integral as usual:

$$EPE(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |y - f(x)|\ pr_{X,Y}(x,y)\ \mathrm{d}x\ \mathrm{d}y \tag{2.37}$$

$$= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} |y - f(x)|\ pr_{Y|X}(y|x)\ \mathrm{d}y\right) pr_X(x)\ \mathrm{d}x. \tag{2.38}$$

It's sufficient to minimize point-wise. Hence, our solution is encoded as follows:

$$f(x) = \underset{c\in\mathbb{R}}{argmin} \int_{-\infty}^{\infty} |y - c|\ pr_{Y|X}(y|x)\ \mathrm{d}y. \tag{2.39}$$

We define a new function $g : \mathbb{R}^2 \to \mathbb{R}$ by

$$g(c,x) = \int_{-\infty}^{\infty} |y - c|\ pr_{Y|X}(y|x)\ \mathrm{d}y \tag{2.40}$$

$$= \int_{-\infty}^{c} (c - y)\ pr_{Y|X}(y|x)\ \mathrm{d}y + \int_{c}^{\infty} (y - c)\ pr_{Y|X}(y|x)\ \mathrm{d}y. \tag{2.41}$$

We compute the partial w.r.t c:

$$\frac{\partial g}{\partial c} = \int_{-\infty}^{c} pr_{Y|X}(y|x)\ \mathrm{d}y - \int_{c}^{\infty} pr_{Y|X}(y|x)\ \mathrm{d}y \tag{2.42}$$

$$= \int_{-\infty}^{c} pr_{Y|X}(y|x) - \left[1 - \int_{-\infty}^{c} pr_{Y|X}(y|x)\right] \tag{2.43}$$

$$= 2\int_{-\infty}^{c} pr_{Y|X}(y|x) - 1 \tag{2.44}$$

Hence, $\frac{\partial g}{\partial c} = 0$ when we have

$$\int_{-\infty}^{c} pr_{Y|X}(y|x) = \frac{1}{2}. \tag{2.45}$$

We note that g is positive definite, hence when c satisfies the above relation, we have a minimizing solution. We note that by definition, the median is the value m satisfying $Pr(Y \leq m|X = x) = 1/2$. Hence,

$$f(x) = median(Y|X = x) \tag{2.46}$$

□

Categorical Variable G

If the output is instead a categorical variable G, then we have to use a different loss function for penalizing prediction errors. We denote $\mathcal{G}$ as the set of all possible classes. Hence, an estimate $\hat{G}$ will assume values in $\mathcal{G}$. Our loss function could be represented by a $K \times K$ matrix $\mathbf{L}$, where $K = card(\mathcal{G})$ (Cardinality of $\mathcal{G}$).

We would therefore construct $\mathbf{L}$ such that $\mathbf{L}(k,l)$ is the penalty for classifying an observation belonging to class $\mathcal{G}_k$ as $\mathcal{G}_l$. Hence, we would want $\mathbf{L}$ to be zero along the diagonal and non-negative elsewhere.

Definition 2.14: Zero-One Loss Function

The zero-one loss function is a loss function where all misclassifications are charged a single unit. In essence, let $\mathbf{L}_{ZO}$ denote the zero-one loss function. Then,

$$\mathbf{L}_{ZO}(k,l) = \delta_{kl} \tag{2.47}$$

Proposition 2.7: Bayes Classifier

Let G denote a categorical variable and $\hat{G}$ denote an estimate. We taken $\mathcal{G}$ to be the set of all possible categorical values, defining $K := card(\mathcal{G})$. Hence, $\mathcal{G}_j$ represents the j^{th} categorical value in the set. Let L denote the loss function for this categorical variable and estimate. The expected prediction error is defined as

$$EPE = E\big[L(G, \hat{G}(X))\big], \tag{2.48}$$

where expectation is taken with respect to the joint distribution $Pr(G, X)$. Hence, we can write

$$EPE = E_X\Bigg[\sum_{k=1}^{K} L(\mathcal{G}_k, \hat{G}(X))Pr(\mathcal{G}_k|X)\Bigg]. \tag{2.49}$$

Subjecting this to pointwise minimization, we have the minimizing solution given by

$$\hat{G}(x) = \underset{g\in\mathcal{G}}{\text{argmin}} \sum_{k=1}^{K} L(\mathcal{G}_k, g)Pr(\mathcal{G}_k|X = x)\big]. \tag{2.50}$$

If we apply the 0-1 loss function, we obtain

$$\hat{G}(x) = \underset{g\in\mathcal{G}}{\text{argmin}} \big[1 - Pr(\mathcal{G}_k|X = x)\big], \tag{2.51}$$

or simply

$$\hat{G}(x) = \mathcal{G}_k \text{ if } Pr(\mathcal{G}_k|X = x) = \max_{g\in\mathcal{G}} Pr(g|X = x). \tag{2.52}$$

This solution is known as the Bayes classifier, which says that we classify to the most probable class, using the conditional (discrete) distribution $Pr(G|X)$.

Definition 2.15: Bayes Rate

The error rate of the Bayes classifier is called the Bayes rate.

2.5 Local Methods in High Dimensions

Curse of Dimensionality

There are many manifestations of this problem, and we will examine a few here

Proposition 2.8: The Median Distance for N data points in p-dimensions

Consider N data points uniformly distributed in a $p-$dimensional unit ball centered at the origin. Suppose that we consider a nearest-neighbour estimate at the origin. The median distance from the origin to the *closest data point is given by the expression*

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p} \tag{2.53}$$

Proof. Let F denote the distribution. We first want to construct the density and distribution for these random variables. We note that the distribution for a point to lie in a concentric sphere of radius r is given by

$$F(r) = \int_{B_r(0)} \mathrm{d}^p x \; f(x), \tag{2.54}$$

where $B_r(0)$ denotes the r-ball centered on the origin and $f(x)$ is the corresponding uniform density satisfying $f(x) = 1/Vol(B_1(0))$. The volume of the unit ball in p-dimensions is given by

$$Vol(B_1(0)) = \int_{B_1(0)} \mathrm{d}^p x = \int \mathrm{d}\Omega_{p-1} \int_0^1 \mathrm{d}x \; x^{p-1} = \frac{\Omega_{p-1}}{p}, \tag{2.55}$$

where Ω_{p-1} denotes the surface area of the unit $p - sphere$. We can therefore compute the probability that a uniformly distributed variable in the p-dimensional unit ball will lie in some concentric sphere of radius r by

$$Pr(X \leq r) = F(r) = \int \mathrm{d}\Omega_{p-1} \int_0^r \mathrm{d}x \; (\frac{p}{\Omega_{p-1}}) x^{p-1} = r^p. \tag{2.56}$$

We are interested in the median distance from the origin to the **closest** data point. To accommodate this, we require the notion of order statistics among N independent identically distributed random variables. In particular, suppose that we have N data points, then let's denote the corresponding ordered random variables by $X_{(1)}, X_{(2)}, ..., X_{(N)}$, with $X_{(j)}$ denoting the j^{th} order statistic. By Proposition A.23, we have the closest statistic distribution

$$F_{(1)}(r) = Pr(X_{(1)} \leq r) = \sum_{k=1}^{N} \binom{N}{k} (1 - F(r))^{N-k} (F(r))^k = 1 - (1 - F(r))^N = 1 - (1 - r^p)^N, \tag{2.57}$$

where the second-last equality is obtained via binomial identity. We note that the median value is by definition the value d that satisfies

$$\frac{1}{2} = F_{(1)}(d). \tag{2.58}$$

Hence, solving for d in

$$\frac{1}{2} = 1 - (1 - d^p)^N, \tag{2.59}$$

we obtain

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p} \tag{2.60}$$

□

Lemma 2.1: Monotonicity of Exponentiation

Let $0 < \alpha < 1$. If $m < n$, then we have that

$$\alpha^{1/m} < \alpha^{1/n}. \tag{2.61}$$

The inequality is flipped if $\alpha > 1$.

Example 2.1: Curse of Dimensionality

By Proposition 2.8, we have a function for the closest median distance for N identical uniformly distributed random variables in the p-ball. Given that $0 < 1 - \frac{1}{2}^{1/N} < 1$ for any $N \in \mathbb{N}$, by Lemma 2.1 we can observe that for a fixed number of data points, they will tend to be closer to the boundary of the p-ball in higher dimensions. That is, if $p < q$, then $d(p, N) < d(q, N)$. This presents a problem as prediction is much more difficult near the edges of the training sample.

Another manifestation of the curse is that the sampling density is proportional to $N^{1/p}$, where p is the dimension of the input space and N is the sample size. Thus, in high dimensions all feasible training samples sparsely populate the input space.

Definition 2.16: Bias

Let $\hat{\nu}$ represent an estimator and ν be a parameter. Then, the bias of our estimator is defined by

$$Bias(\hat{\nu}) = E[\hat{\nu}] - \nu \tag{2.62}$$

This quantity is also referred to as the long-run average error of $\hat{\nu}$.

Proposition 2.9: Bias-Variance Decomposition

Let $\mathcal{T}$ be a training data set and $\hat{\theta}$ be an estimator for the parameter θ. Then, we define the mean squared error as $MSE(\theta, \hat{\theta}) = E_{\mathcal{T}}\big[(\hat{\theta} - \theta)^2\big]$. We therefore have the identity

$$MSE(\theta, \hat{\theta}) = Var_{\mathcal{T}}[\hat{\theta}] + \big(Bias(\hat{\theta})\big)^2, \tag{2.63}$$

known as the Bias-Variance Decomposition.

Proof. We compute

$$\begin{aligned} MSE(x_0) &= E_{\mathcal{T}}\big[f(x_0) - \hat{y}_0\big]^2 & (2.64)\\ &= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)]^2 + \big[E_{\mathcal{T}}(\hat{y}_0) - f(x_0)\big]^2 & (2.65)\\ &= Var_{\mathcal{T}}(\hat{y}_0) + Bias^2(\hat{y}_0) & (2.66)\end{aligned}$$

□

2.6 Statistical Models, Supervised Learning and Function Approximation

Failure of Nearest-Neighbours

Our primary goal is to find a useful approximation $\hat{f}(x)$ to the function f(x) that underlies the predictive relationship between the inputs and outputs. The class of nearest-neighbor methods can be viewed as direct estimates of the regression function seen in §2.4: $f(x) = E(Y|X = x)$. However, we have seen that they can fail in at least two ways:

- *If dimension of input space is high, the nearest neighbors need not be close to the target point, and can result in large errors;*

• *If special structure is known to exist, this can be used to reduce both the bias and variance of the estimates.*

2.6.1 A Statistical Model for the Joint Distribution Pr(X,Y)

Proposition 2.10

Suppose that our data arose from a statistical model

$$Y = f(X) + \epsilon, \tag{2.67}$$

where the random error ϵ has $E(\epsilon) = 0$ and is independent of X. We note that for this model, using the squared error loss function gives us the solution $\hat{f}(x) = E\big[Y|X = x\big]$.

Proof. We compute the squared error loss:

$$\begin{aligned} EPE(f) = E[(Y - \hat{f}(X) - \epsilon)^2] &= E[(Y - \hat{f}(X))^2 - 2\epsilon(Y - f(X)) + \epsilon^2] && (2.68)\\ &= E[(Y - \hat{f}(X))^2] - 2E[\epsilon]E[Y - \hat{f}(X)] + E[\epsilon^2] && (2.69)\\ &= E[(Y - \hat{f}(X))^2] + Var[\epsilon] && (2.70)\end{aligned}$$

The minimizing solution to this was shown in Proposition 2.3. We note that the presence of $Var[\epsilon]$ is irrelevant as it serves as background noise, that is not dependent on f(x). Hence, we have our solution

$$\hat{f}(x) = E[Y|X = x] \tag{2.71}$$

□

This additive error model is a useful approximation to the truth. For most system, the input-output pairs (X,Y) will not have a deterministic relationship $Y = f(X)$. There will generally be other unmeasured variables that also contribute to Y, such as measurement error. The additive model therefore assumes that we can capture all these departures from a deterministic relationship through the error ϵ.

The assumption in (2.67) is that errors are independent and identically distributed. These assumptions are not strictly necessary but seems like a good model when we average squared errors uniformly in our EPE criterion.

2.6.2 Supervised Learning

We want to present the function-fitting paradigm from a machine learning point of view. Suppose for simplicity that the errors are additive and that the model $Y = f(X) + \epsilon$ is a reasonable assumption. Supervised learning aims to learn f by example through a *teacher*. One assembles a *training* set of observations $\mathcal{T} = (x_i, y_i)$, $i = 1, ..., N$ for the inputs and outputs in the system of study. There is a learning algorithm that is fed observed input values x_i which produces outputs $\hat{f}(x_i)$ in response to the inputs. The learning algorithm has the property that it can modify its input/output relationship $\hat{f}$ in response to differences $y_i - \hat{f}(x_i)$ between original and generated outputs. This process is known as *learning by example*. Upon completion of the learning process, the hope is that artificial and real outputs are sufficiently close enough to be useful for all sets of inputs likely encountered in practice.

2.6.3 Function Approximation

Learning paradigm of previous section has been the motivation for research into supervised learning problems in the field of machine learning (with analogies to human reasoning) and neural networks (with biological analogies to the brain). One can consider the data pairs $\{x_i, y_i\}$ as points in (p+1)-dimensional Euclidean space. The function f(x) has domain equal to the p-dimensional input subspace, and is related to the data via a model such as $y_i = f(x_i) + \epsilon_i$. For convenience, we will assume the domain to be $\mathbb{R}^p$. The goal is to obtain a useful approximation to f(x) for all x in some region of $\mathbb{R}^p$, given the representations in $\mathcal{T}$.

Instead of the aforementioned learning paradigm, we will treat supervised learning as a problem in function approximation. This encourages the geometrical concepts of Euclidean spaces and mathematical concepts of probabilistic inference to be applied to the problem. This will be the approach taken in this book.

Definition 2.17: Linear Basis Expansions

Many of the approximation that we will encounter have an associated set of parameters θ that can be modified to suit the data at hand. For instance, the linear model $f(x) = x^T\beta$ has $\theta = \beta$. Another class of useful approximators can be expressed as linear basis expansions

$$f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k, \tag{2.72}$$

where the h_k are a suitable set of functions or transformations of the input vector x. Some traditional examples for h_k are polynomials and trigonometric expressions. We'll also encounter nonlinear expansions, such as sigmoid transformation common to neural network models,

$$h_k(x) = \frac{1}{1 + exp(-x^T\beta_k)} \tag{2.73}$$

Maximum Likelihood Estimation

Suppose that we had N random variables associated with N measurements that we denote by $X_1, .., X_N$. Then, we denote their joint density, parameterized by θ by

$$f_\theta(x_1, ..., x_N) = f(x_1, ..., x_N|\theta), \tag{2.74}$$

where we have observed values $X_1 = x_1, ..., X_N = x_N$. We define the likelihood of θ as the function

$$lik(\theta) = f(x_1, ..., x_N|\theta). \tag{2.75}$$

If the distribution is discrete, then f is the frequency distribution function. In essence, $lik(\theta)$ measures the probability that the observed data was generated by a distribution parameterized by θ. Hence, the principle of maximum likelihood estimation argues that one should fit the model by the θ such that $lik(\theta)$ is maximized. Formally, we want

$$\hat{\theta} = \underset{\theta \in S}{argmax}\ lik(\theta), \tag{2.76}$$

where S denotes the set of all possible values for θ.

Definition 2.18: Residual Sum of Squares, RSS

Suppose that we have a set of N data points (i.e pairing of observed input-outputs (x, y)). This set would be given by $\{(x_i, y_i)|i \in \mathbb{Z}_N\}$. Then, we can estimate the parameters θ in f_θ as we did for the linear model *via minimizing residual sum-of-squares:*

$$RSS(\theta) = \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2 \tag{2.77}$$

Example 2.2: Max Likelihood Estimation: Least Squares

Suppose that we have a random sample y_i, $i = 1, ..., N$ from a density $Pr_\theta(y)$ indexed by some parameters θ. Then, the likelihood function is given by

$$lik(\theta) = Pr(y_1, ..., y_N|\theta) = \prod_{i=1}^{N} Pr(y_i;\theta), \tag{2.78}$$

where the second equality has assumed that y_i's are independent. We note that although likelihood is often stated in a conditional formalism $Pr(Y|\theta)$, θ is not a random variable but an unknown parameter. Hence, we will typically write $Pr(Y;\theta)$. Then, the log-probability of the observed sample is given by

$$L(\theta) = \sum_{i=1}^{N} \log[Pr(y_i;\theta)]. \tag{2.79}$$

Since log is a monotonically increasing function, it preserves the solution to MLE:

$$\hat{\theta} = \underset{\theta\in S}{argmax} \prod_{i=1}^{N} Pr(y_i;\theta) = \underset{\theta\in S}{argmax} \sum_{i=1}^{N} \log[Pr(y_i;\theta)]. \tag{2.80}$$

Then, suppose that we consider the additive error model $Y = f_\theta(X) + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$. Then, suppose that the conditional likelihood was Gaussian

$$Pr(Y|X,\theta) = N(f_\theta(X), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y-f_\theta(X))^2}{2\sigma^2}} \tag{2.81}$$

We therefore have

$$L(\theta) = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-f_\theta(x_i))^2}{2\sigma^2}}\right] \tag{2.82}$$

$$= -\frac{N}{2} log(2\pi) - N log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2. \tag{2.83}$$

Since only the last term contains θ, by MLE we have that

$$\hat{\theta} = \underset{\theta\in S}{argmax}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2\right) = \underset{\theta\in S}{argmin}\ \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 = \underset{\theta\in S}{argmin}\ RSS(\theta), \tag{2.84}$$

which tells us that our solution to MLE is found by minimizing $RSS(\theta)$. We note that this is typically how loss functions can be derived. We choose some reasonable distribution that may have generated the data, invoke the principle of MLE and derive the function that requires minimization (i.e a loss function).

Example 2.3: Multinomial Likelihood

Consider the multinomial likelihood for the regression function $Pr(G|X)$ for a qualitative output G. Suppose that we have a model $Pr(G = \mathcal{G}_k|X = x) = p_{k,\theta}(x)$, $k = 1..., K$ for the conditional probability of each class given X, indexed by parameter vector θ. Then, the log-likelihood (also referred to as cross-entropy) is

$$L(\theta) = \sum_{i=1}^{N} log(p_{g_i,\theta}(x_i)), \tag{2.85}$$

and when maximized it delivers values of θ that best conforms with the data in the likelihood sense.

2.7 Structured Regression Models

We've seen that although nearest-neighbour and other local methods focus directly on estimating the function at a point, they face problems in high dimensions. However, they may also be inappropriate in low dimensions where more structured approaches can make more efficient use of the data. We'll introduce classes of such structured approaches in this section.

2.7.1 Difficulty of the Problem

Consider the RSS criterion for an arbitrary function f,

$$RSS(f) = \sum_{i=1}^{N}(y_i - f(x_i))^2 \tag{2.86}$$

Note that minimizing (2.86) leads to infinitely many solutions, as any function $\hat{f}$ passing through the training points (x_i, y_i) is a solution. One would also run into the problem of overfitting as any particular solution chosen might be a poor predictor at test points different from the training points. If there are multiple observation pairs x_i, y_{il}, $l = 1, ..., N_i$ at each value of x_i, the risk becomes limited as the solutions would pass through average values of the y_{il} at each x_i.

In order to obtain useful results for finite N, we must restrict the eligible solutions to (2.86) to a smaller set of functions. Any restrictions imposed on f that lead to a unique solution to (2.86) does not really remove the ambiguity caused by the vast space of solutions. There are infinitely many possible restrictions, each leading to the unique solution.

In general, the imposed constraints by most learning methods can be described as *complexity* restrictions of some kind. Usually, this means some kind of regular behaviour in small neighbourhoods of the input space. That is, for all input points x sufficiently close to each other in some metric, $\hat{f}$ exhibits some special structure such as nearly constant, linear or low-order polynomial behaviour. The estimator can then be obtained by averaging or polynomial fitting in that neighbourhood.

Methods such as splines, neural networks and basis-function methods implicitly define neighbourhoods of local behaviour. Any method that attempts to produce locally varying functions in small isotropic neighbourhoods will run into problems in high dimensions - curse of dimensionality. In addition, all method that overcome dimensionality problems have an associated - and often implicit or adaptive- metric for measuring neighbourhoods, which basically does not allow the neighbourhood to be simultaneously small in all directions.

2.8 Classes of Restricted Estimators

Here we will give a brief summary, since detailed descriptions are given in later chapters. Each of the classes has associated with it one or more parameters, sometimes appropriately called *smoothing* parameters.

2.8.1 Roughness Penalty and Bayesian Methods

Here, the class of functions is controlled by explicitly penalizing $RSS(f)$ with a roughness penalty

$$PRSS(f;\lambda) = RSS(f) + \lambda J(f). \tag{2.87}$$

The user-selected functional $J(f)$ will be large for functions f that vary too rapidly over small regions of input space.

Example 2.4: Cubic Smoothing Spline

For example, the popular cubic smoothing spline *for one-dimensional inputs is the solution to the penalized least-squares criterion*

$$PRSS(f;\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx \tag{2.88}$$

The roughness penalty here controls large values of the second derivative of f, and the amount of penalty is dictated by $\lambda \geq 0$*. For* $\lambda = 0$ *no penalty is imposed, and any interpolating function will do, while for* $\lambda = \infty$ *only functions linear in x are permitted. We can observe this fact by recognizing that the second term must go to zero if* $\lambda \to \infty$*. For this to occur, we require that* $f''(x) = 0$*, which are the class of functions* $f(x) = \alpha x + b$ *for some* $\alpha, b \in \mathbb{R}$*.*

Example 2.5: Additive Functionals

Penalty functionals J can be constructed for functions in any dimension, and special versions can be created to impose special structure. For example, additive penalties $J(f) = \sum_{j=1}^{p} J(f_j)$ *are used in conjunction with additive functions* $f(X) = \sum_{j=1}^{p} f_j(X_j)$ *to create additive models with smooth coordinate functions.*

Example 2.6: Projection Pursuit Regression

Projection pursuit regression models have $f(X) = \sum_{m=1}^{M} g_m(\alpha_m^T X)$ *for adaptively chosen directions* α_m*, and the functions* g_m *can each have an associated roughness penalty.*

A penalty function, or *regularization methods*, express our prior belief that type of functions we seek exhibit a certain type of smooth behaviour, and can be cast in a Bayesian framework. The penalty J corresponds to a log-prior, and $PRSS(f;\lambda)$ the log-posterior distribution, and minimizing $PRSS(f;\lambda)$ amounts to finding the posterior mode.

2.8.2 Kernel Methods and Local Regression

These methods aim to explicitly provide estimates of the regression function or conditional expectation by specifying the nature of the local neighbourhood, and of class of regular functions fitted locally.

Definition 2.19: Kernel Density Estimation

In statistics, kernel density estimation *(KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.*

Definition 2.20: Kernel Function

The local neighbourhood is specified by a kernel function $K_\lambda(x_0, x)$ which assigns weights to points x in a region around x_0. For example, the Gaussian kernel has a weight function based on the Gaussian density function

$$K_\lambda(x_0, x) = \frac{1}{\lambda} exp\left[- \frac{||x - x_0||^2}{2\lambda} \right], \tag{2.89}$$

and assigns weights to points that die exponentially with their squared Euclidean distance from x_0. The parameter λ corresponds to the variance of Gaussian density, and controls the width of the neighbourhood.

Definition 2.21: Nadaraya-Watson Weighted Average

The simplest form of kernel estimate is the Nadaraya-Watson weighted average

$$\hat{f}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^N K_\lambda(x_0, x_i)} \tag{2.90}$$

Definition 2.22: Local Regression Estimate

In general, we can define a local regression estimate of $f(x_0)$ as $f_{\hat{\theta}}(x_0)$, where $\hat{\theta}$ minimizes

$$RSS(f_\theta, x_0) = \sum_{i=1}^N K_\lambda(x_0, x_i)(y_i - f_\theta(x_i))^2, \tag{2.91}$$

and f_θ is some parameterized functions, such as a low-order polynomials.

Example 2.7: Parameterized Functions for Local Regression Estimate

Some examples of parameterized functions for local regression estimate are

- *$f_\theta(x) = \theta_0$, the constant function. This results in the Nadaraya-Watson estimate mentioned above in (2.90).*
- *$f_\theta(x) = \theta_0 + \theta_1 x$ gives the popular local linear regression model.*

Nearest-neighbour methods can be though of as kernel methods having a more data-dependent metric. Indeed, the metric for k-nearest neighbours is

$$K_k(x, x_0) = I(||x - x_0|| \leq ||x_{(k)} - x_0||), \tag{2.92}$$

where $x_{(k)}$ is the training observation ranked k^{th} in distance from x_0, and I(S) is the indicator of the set S.

2.8.3 Basis Functions and Dictionary Methods

Definition 2.23: Basis Functions

This class of methods includes the familiar linear and polynomial expansions, but more importantly a wide *variety of more flexible models. The model for f is a linear expansion of basis functions*

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x), \tag{2.93}$$

where each of the h_m is a function of the input x, and term linear here refers to the action of the parameters θ. This class covers wide variety of methods.

Definition 2.24: Radial Basis Functions

Radial basis functions are symmetric p-dimensional kernels located at particular centroids,

$$f_\theta(x) = \sum_{m=1}^{M} K_{\lambda_m}(\mu_m, x)\theta_m; \tag{2.94}$$

for instance, the Gaussian kernel $K_\lambda(\mu, x) = e^{-||x-\mu||^2/2\lambda}$ is popular. Radial basis functions have centroids μ_m and scales λ_m that have to be determined.

Definition 2.25: Dictionary Methods

A single-layer feed-forward neural network model with linear output weights can be though of as an adaptive basis function method. The model has the form

$$f_\theta(x) = \sum_{m=1}^{M} \beta_m \sigma(\alpha_m^T x + b_m), \tag{2.95}$$

where $\sigma(x) = 1/(1+e^{-x})$ is known as the activation function. The directions α_m and bias terms b_m have to be determined. These adaptively chosen basis function methods are also known as dictionary methods.

2.9 Model Selection and the Bias-Variance Tradeoff

All models described above and many others that will be discussed have a *smoothing* or *complexity* parameter that has to be determined:

- the multiplier of the penalty term;
- the width of the kernel;
- or the number of basis functions.

Proposition 2.11: Test / Generalization Error

Let $\hat{f}_k(x_0)$ denote the k-nearest neighbour regression fit. The consideration of the nearest neighbours usefully illustrates the competing forces that affect the predictive ability of such approximations. Suppose that data arises from a model $Y = f(X) + \epsilon$, with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. For simplicity, we assume that values of x_i in sample are fixed in advance (non-random). The expected prediction error at x_0, also known as test *or generalization error, can be decomposed:*

$$EPE_k(x_0) = E\big[(Y - \hat{f}_k(x_0)))^2 | X = x_0\big] = \sigma^2 + \big[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\big]^2 + \frac{\sigma^2}{k} \tag{2.96}$$

Proof. We'll compute this by way of the property of conditional expectations [See Proposition A.24]

$$\begin{aligned}
E\big[(Y - \hat{f}_k(x_0)))^2 | X = x_0\big] &= E\big[(\epsilon + f(X) - \hat{f}_k(x_0)))^2 | X = x_0\big] \\
&= E\big[\epsilon^2 + 2\epsilon(f(X) - \hat{f}_k(x_0)) + (f(X) - \hat{f}_k(x_0))^2 | X = x_0\big] \\
&= E\big[\epsilon^2 | X = x_0] + 2E[\epsilon(f(X) - \hat{f}_k(x_0)) | X = x_0] + E[(f(X) - \hat{f}_k(x_0))^2 | X = x_0] \\
&= E[\epsilon^2] + 2E[\epsilon]E[(f(X) - \hat{f}_k(x_0)) | X = x_0] + (f(x_0) - \hat{f}_k(x_0))^2 \\
&= \sigma^2 + \big[Bias^2(\hat{f}_k(x_0)) + Var_{\mathcal{T}}(\hat{f}_k(x_0))\big] \\
&= \sigma^2 + \big(f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\big)^2 + \frac{\sigma^2}{k}, \qquad (2.97)
\end{aligned}$$

where we refer to the first term σ^2 as the *irreducible error*. This is the variance of the new test target-and is beyond our control, even if we know the true $f(x_0)$. The bias term is the squared difference between the true mean $f(x_0)$ and expected value of the estimate $[E_{\mathcal{T}}(\hat{f}_k(x_0)) - f(x_0)]^2$, where expectation averages randomness in training data. We expect that the second term will increase as k increases. We note that choosing larger number of nearest neighbours will influence the average approximation, which should begin to further deviate from $f(x_0)$ as k increases. □

Lemma 2.2: Nearest Neighbours Bias

We consider the case under which the Bias increases as the number of nearest neighbours increases. We fix a point $x_0 \in \mathbb{R}^p$. Let (x_i, y_i) denote an input-output pair from a training sample of size m. We define $x_{(i)}$ $(1 \leq i \leq m)$ as the ordered set of points that are closest to x_0. Let N_k denote the k-nearest neighbours estimate

$$N_k = \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}), \tag{2.98}$$

where $f(x_{(i)}) = y_i$. The Bias term

$$Bias(k) = f(x_0) - N_k \tag{2.99}$$

increases as k increases if

$$N_k > \frac{1}{n-k}\sum_{i=k+1}^{n} f(x_{(i)}) \tag{2.100}$$

where $n > k$. The term on the right hand side of the inequality measures the average y_i value of the k+1th to n^{th} farthest away neighbour from x_0. Hence, if the average of farther away neighbours tends to be less than k-nearest neighbours, the Bias will increase.

Proof. We want to observe the conditions under which

$$f(x_0) - N_k < f(x_0) - N_n \tag{2.101}$$

From this inequality, we have that

$$N_k > N_n \tag{2.102}$$

We can relate N_n to N_k via

$$N_n = \frac{k}{n}N_k + \frac{1}{n}\sum_{i=k+1}^{n} f(x_{(i)}). \tag{2.103}$$

Then, we have that

$$\frac{k}{n}N_k + \frac{1}{n}\sum_{i=k+1}^{n} f(x_{(i)}) < N_k \tag{2.104}$$

$$\to N_k > \frac{1}{n-k}\sum_{i=k+1}^{n} f(x_{(i)}) \tag{2.105}$$

□

Bias-Variance Tradeoff

The bias-variance tradeoff is a central problem in supervised learning. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with high bias typically produce simpler models that don't tend to overfit but may underfit their training data, failing to capture important regularities. The variance term in (2.97) decreases as the inverse of k. Hence, as k varies, there is a bias-variance tradeoff.

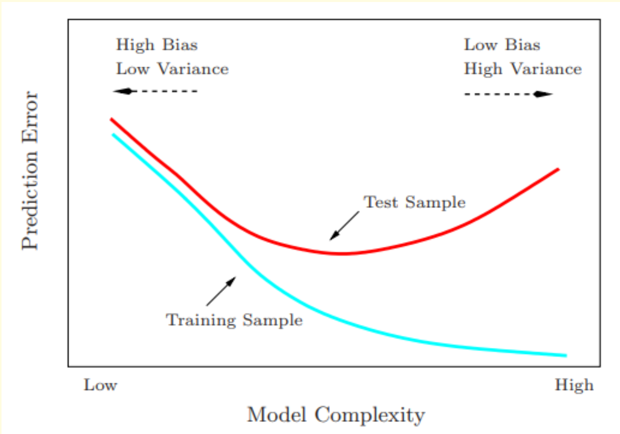

Figure 2.1: Test and training error as a function of model complexity.

Typically, we would like to choose our model complexity to trade bias off with variance in such a way as to minimize the test error. An obvious estimate of test error is the training error $\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$*. However, training error is not a good estimate of test error, as it does not properly account for model complexity. Figure 2.1 shows typical behaviour of test and training error, as model complexity is varied. If we perform* *too much fitting, the model adapts itself too closely to the training data, and will not generalize well.*

3 CH.3: Linear Methods for Regression

3.1 Introduction

A linear regression model assumes that the regression function $E(Y|X)$ is linear in the inputs $X_1, ..., X_p$. For prediction purposes, linear models can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data. In this chapter, we describe linear methods for regression. The authors believe that an understanding of linear methods is essential for understanding nonlinear ones. Many nonlinear techniques can be considered direct generalizations of the linear methods that we'll discuss here.

3.2 Linear Regression Models and Least Squares

Linear Model

Just as in §2, we have an input vector $X^T = (X_1, X_2, ..., X_p)$, and want to predict a real-valued output Y. The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{3.1}$$

The linear model either assumes that the regression function $E(Y|X)$ is linear, or that the linear model is a reasonable approximation. Here the β_j's are unknown parameters or coefficients, and the variables X_j can come from a variety of sources:

- *quantitative inputs;*
- *transformations of quantitative inputs, such as log, square-root or square;*
- *basis expansions, such as $X_2 = X_1^2$, $X_3 = X_1^3$, leading to polynomial representation;*
- *interactions between variables, such as $X_3 = X_1 \cdot X_2$*

Proposition 3.1

Let X be a $n \times m$ matrix. If X has full column rank, then $X^T X$ is positive definite.

Proof. Let $x_1, ..., x_n$ denote the column vectors of X. Then, we can express X in terms of these column vectors as

$$X = \begin{bmatrix} | & | & ... & | \\ | & | & ... & | \\ x_1 & x_2 & ... & x_n \\ | & | & ... & | \\ | & | & ... & | \end{bmatrix}, \tag{3.2}$$

If X has full column rank, then all of its columns are linearly independent. Let $v_i \in \mathbb{R}\ \forall i$, then the only solution to

$$\sum_{i=1}^{n} x_i v_i = \mathbf{0} \tag{3.3}$$

is if $v_i = 0\ \forall i$. Let $v \in \mathbb{R}^n$. As a consequence of this, $Xv = \mathbf{0}$ iff $v = \mathbf{0}$. We will now show this. We can observe that

$$Xv = \sum_{i=1}^{n} x_i v_i, \tag{3.4}$$

which by its full column rank property means that $Xv \neq 0\ \forall v \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. Let $y = Xv$. Then, we consequently have

$$y^T y = v^T X^T X v = \sum_{i=1}^{m} y_i^2 > 0, \tag{3.5}$$

which by the definition of positive definite matrices, means that $X^T X$ is positive definite. □

Corollary 3.1

Let A be a $n \times n$ matrix. If A is positive definite, then A is invertible.

Proof. Since A is positive definite, then $v^T A v > 0\ \forall v \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. Hence, $Av \neq \mathbf{0}\ \forall v \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. Consequently, A must have full column rank. Since A has full column rank, A is invertible. □

Least Squares Estimation

Suppose that we have a set of training data $(x_1, y_1), ..., (x_N, y_N)$ from which to estimate the parameters β. Each $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ is a vector of feature measurements from the i^{th} case. Most popular estimation method is least squares, in which we picks coefficients $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ to minimize residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \tag{3.6}$$

From a statistical point of view the criterion is reasonable if the y_i's are conditionally independent given the inputs x_i. We define $\mathbf{X}$ as the $N \times (p+1)$ matrix with each row being an input vector. Similarly, let $\mathbf{y}$ be the N-vector of outputs in the training set. Then, we can write the residual sum-of-squares as

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta). \tag{3.7}$$

Let $Y := \mathbf{y} - \mathbf{X}\beta$. Differentiating w.r.t β gets us

$$\begin{aligned}\frac{\partial RSS}{\partial \beta} &= \frac{\partial Y}{\partial \beta}(\mathbf{y} - \mathbf{X}\beta) + (\mathbf{y} - \mathbf{X}\beta)^T \frac{\partial Y}{\partial \beta} && (3.8)\\ &= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) + (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{X} && (3.9)\\ &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) && (3.10)\end{aligned}$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X} \quad (3.11)$$

Suppose that $\mathbf{X}$ has full column rank, then by Proposition 3.1 and Corollary 3.1, X^TX is positive definite and is invertible. Setting first derivative to zero:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0, \quad (3.12)$$

has unique solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (3.13)$$

The fitted values at the training inputs are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}, \quad (3.14)$$

where $\hat{y}_i = \hat{f}(x_i)$. The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})\mathbf{X}^T$ appearing in equation (3.14) is sometimes referred to as the hat matrix because it puts the hat on $\mathbf{y}$.

Geometrical Interpretation

A Probability Theory

A.1 Constructing a Probability Space

The first order of things is to establish the fundamental structures that comprise a probability space. We will first label a set Ω such that its points ω are associated with possible outcomes of a measurement. We also denote $\mathcal{A}$ to be a nonempty collection of subsets of Ω which will represent collection of *events* that will be assigned probabilities.

Definition A.1: Sample Space, Ω

A set Ω with outcomes $s_1, s_2, ..s_n$ (i.e. $\Omega = \{s_1, s_2, \dots, s_n\}$) must meet some conditions in order to be a sample space:
• The outcomes must be mutually exclusive, i.e. if s_j takes place, then no other s_i will take place, $\forall i, j \in \{1, 2, \dots, n\} \quad i \neq j$.
• The outcomes must be collectively exhaustive, i.e., on every experiment (or random trial) there will always take place some outcome $s_i \in \Omega$ for $i \in \{1, 2, \dots, n\}$.
• The sample space Ω must have the right granularity depending on what we are interested in. We must remove irrelevant information from the sample space. In other words, we must choose the right abstraction (forget some irrelevant information).

Definition A.2: σ-algebra / σ-field $\mathcal{A}$ [Event Space]

A non-empty collection of subsets $\mathcal{A}$ of set Ω is called a σ-field of subsets of Ω provided that the following two properties hold:

(i) If A is in $\mathcal{A}$, then A^c is also in $\mathcal{A}$.
(ii) If A_n is in $\mathcal{A}$, $n = 1, 2, ...$, then $\cup_{n=1}^{\infty} A_n$ and $\cap_{n=1}^{\infty} A_n$ are both in $\mathcal{A}$.

Definition A.3: Event

Given a σ-field $\mathcal{A}$ that corresponds to some sample space Ω. We say that if $A \in \mathcal{A}$, then A is an event.

The statement 'the event A occurs' means that the outcome of our experiment is represented by some point $\omega \in A$. For an event A, if we let $P(A)$ denote the probability of the event, then we have $0 \leq P(A) \leq 1$.

Definition A.4: Probability Measure

A probability measure P on a σ-field of subsets $\mathcal{A}$ of a set Ω is a real valued function having domain $\mathcal{A}$ satisfying the following properties:

(i) $P(\Omega) = 1$
(ii) $P(A) \geq 0 \;\; \forall A \in \mathcal{A}$
(iii) If A_n, $n = 1, 2, 3, ...$ are mutually disjoint sets in $\mathcal{A}$, then $P\big(\cup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} P(A_n)$.

Definition A.5: Probability Space

A probability space, denoted by $(\Omega, \mathcal{A}, P)$ is a set Ω, a σ-field of subsets $\mathcal{A}$, and probability measure P defined on $\mathcal{A}$.

A.2 Standard Definitions and Properties

Definition A.6: Conditional Probability

Let A and B be two events such that $P(A) > 0$. Then the conditional probability of B given A, written $P(B|A)$, is defined to be

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{A.1}$$

If $P(A) = 0$, then the conditional probability of B given A is undefined.

Proposition A.1

Let A be an event and A^c be its complement, defined as $A^c = \Omega - A$. It follows from the properties of disjoint probability sets that

$$P(A^c) = 1 - P(A) \tag{A.2}$$

Definition A.7: Independent Events

Two events A and B are independent if and only if

$$P(A \cap B) = P(A)P(B) \tag{A.3}$$

This definition emerges as a consequence of wanting to construct a notion of an event's occurrence having no influence on the the occurrence of the other event. Through the conditional probabilistic lens, this would mean $P(B|A) = P(B)$ (i.e. Given that A has occurred, this does not affect the probability that B will occur). Therefore, it follows that $P(A \cap B) = P(A)P(B)$.

Definition A.8: Mutual Exclusivity

Events A and B are said to be two mutually exclusive events if both cannot occur. In essence, their intersection is disjoint $A \cap B = \emptyset$ so that they have the following properties:

$$P(A \cap B) = 0 \tag{A.4}$$

$$P(A \cup B) = P(A) + P(B) \tag{A.5}$$

Definition A.9: Discrete Random Variable

A discrete real-valued random variable X on a probability space $(\Omega, \mathcal{A}, P)$ is a function X with domain Ω and range that is a finite or countably infinite subset $\{x_1, x_2, ...\}$ of the real numbers $\mathbb{R}$ such that $\{\omega : X(\omega) = x_i\}$ is event for all i.

Hence, $\{\omega : X(\omega) = x_i\}$ is an event and we usually will write $\{X = x_i\}$ for brevity and denote the probability of this event as $P(X = x_i)$.

Definition A.10: Discrete Density Function

The real-valued function f defined on $\mathbb{R}$ by $f(x) = P(X = x)$ is called the discrete density function of X. A number x is called a possible value of X if $f(x) > 0$.

We note that a real-valued function f defined on $\mathbb{R}$ is called a discrete density function provided that it satisfies the following properties:

(i) $f(x) \geq 0$, $x \in \mathbb{R}$.
(ii) $\{x : f(x) \neq 0\}$ *is a finite or countably infinite subset of* $\mathbb{R}$. *Let* $\{x_1, x_2, ...\}$ *denote this set. Then*
(iii) $\sum_i f(x_i) = 1$.

We can compute the probability of X taking on value in some set A via

$$P(X \in A) = \sum_{x \in A} f(x) \tag{A.6}$$

Definition A.11: Discrete r-dimensional Random Vector

We let $\mathbb{R}^r$ denote the collection of all r-tuples of real numbers. A point $\mathbf{x} = (x_1, x_2, ..., x_r)$ of $\mathbb{R}^r$ is usually called an r-dimensional vector. Thus for each $\omega \in \Omega$, the r values $X_1(\omega), .., X_r(\omega)$ define a point

$$X(\omega) = (X_1(\omega), X_2(\omega), ..., X_r(\omega)) \tag{A.7}$$

of $\mathbb{R}^r$. This defines an r-dimensional vector-valued function on Ω, $\mathbf{X} : \Omega \to \mathbb{R}^r$, which is usually written as $\mathbf{X} = (X_1, X_2, ..., X_r)$.

A discrete r-dimensional random vector $\mathbf{X}$ is a function $\mathbf{X}$ from Ω tp $\mathbb{R}^r$ taking on a finite or countably infinite number of values $\mathbf{x}_1, \mathbf{x}_2, ..$ such that

$$\{\omega : \mathbf{X}(\omega) = x_0\} \tag{A.8}$$

is an event for all i.

Definition A.12: Discrete Density Function for Random Vector

The discrete density function f for the random vector $\mathbf{X}$ is defined by

$$f(x_1, ..., x_r) = P(X_1 = x_1, ..., X_r = x_r) \tag{A.9}$$

or equivalently

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^r \tag{A.10}$$

The probability that $\mathbf{X}$ belongs to the subset A of $\mathbb{R}^r$ can be found by using the analog of (A.6), namely

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}) \tag{A.11}$$

Definition A.13: Mutually Independent Random Variables

Let $X_1, X_2, ..., X_r$ be r discrete random variables having densities $f_1, f_2, ..., f_r$ respectively. These random variables are said to be mutually independent *if their joint density function f is given by*

$$f(x_1, x_2, ..., x_r) = f_1(x_1) f_2(x_2) \cdots f_r(x_r) \tag{A.12}$$

Consider two independent discrete random variables having densities f_X and f_Y, respectively, Then for any two subsets A and B of R, we have

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B) \tag{A.13}$$

Definition A.14: Probability Generating Function

Let X be a non-negative integer-valued random variable. The probability generating function Φ_X of X is defined as

$$\Phi_X(t) = \sum_{x=0}^{\infty} P(X = x)t^x = \sum_{x=0}^{\infty} f_X(x)t^x, \qquad -1 \leq t \leq 1 \tag{A.14}$$

Definition A.15: Random Variable

A random variable X on a probability space $(\Omega, \mathcal{A}, P)$ is a real-valued function $X(\omega)$, $\omega \in \Omega$, such that for $-\infty < x < \infty$, $\{\omega | X(\omega) \leq x\}$ is an event.

Definition A.16: Continuous Random Variable

A random variable X is called a continuous random variable if

$$P(X = x) = 0, \qquad -\infty < x < \infty \tag{A.15}$$

We can observe that X is a continuous random variable if and only if its distribution function F is continuous at every x, that is, F is a continuous function.

Definition A.17: Symmetric Random Variable

A random variable X is said to be symmetric if X and -X have the same distribution function.

Definition A.18: Median

For any probability distribution on the real line $\mathbb{R}$ with cumulative distribution function F, regardless of whether it is any kind of continuous probability distribution, in particular an absolutely continuous distribution, or a discrete probability distribution, a median is by definition any real number m that satisfies the inequalities

$$P(X \leq m) = \frac{1}{2}, \qquad P(X \geq m) = \frac{1}{2} \tag{A.16}$$

A.3 Distributions and Densities

Let X and Y be two discrete random variables. For any real numbers x and y, the set $\{\omega | X(\omega) = x and Y(\omega) = y\}$ is an event that we will usually denote by $\{X = x, Y = y\}$.

Definition A.19: Joint Density and Marginal Density

Let $\mathbf{X} = (X_1, X_2, ..., X_r)$ be an $\mathbf{r}$-dimensional random vector with density f. Then the function f is usually called the joint density *of the random variables $X_1, X_2, ..., X_r$. The density function of the random variable X_i is then called the* ith marginal density *of $\mathbf{X}$ or of f.*

Definition A.20: (Cumulative) Distribution Function [Discrete]

The function $F(t)$, $-\infty < t < \infty$, defined by

$$F(t) = P(X \leq t) = \sum_{x \leq t} f(x), \qquad -\infty < t < \infty \tag{A.17}$$

is called the distribution function of the random variable X or of the density f. One immediate consequence of this is that it satisfies:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{A.18}$$

Proposition A.2

Let X and Y be independent, non-negative integer-valued random variables. Then

$$\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t) \tag{A.19}$$

Definition A.21: (Cumulative) Distribution Function [Continuous]

The distribution function F of a random variable X is the function

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty \tag{A.20}$$

Proposition A.3: Properties of Distribution Functions

Not all functions can arise as distribution functions, for the latter must satisfy certain conditions. Let X be a random variable and let F be its distribution function. Then

(i) $0 \leq F(x) \leq 1$ *for all x.*
(ii) F *is a non-decreasing function of x.*
(iii) $F(-\infty) = 0$ *and* $F(+\infty) = 1$.
(iv) F(x+) = F(x) for all x. (F is a right-continuous function)

We note that a distribution function is any function F satisfying properties (i)-(iv).

Definition A.22: Probability Density Function (PDF) / Density

A density function / PDF (with respect to integration) is a non-negative function f such that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{A.21}$$

Note that if f is density function, then the function F defined by

$$F(x) = \int_{-\infty}^{x} f(y)\, dy, \qquad -\infty < x < \infty \tag{A.22}$$

*is a continuous function satisfying properties (i)-(iv) in **Prop** A.3.*

Definition A.23: Uniform Density

Let Ω be a sample space with finite measure $Vol(\Omega) < \infty$. Then, a uniform density is a constant function f, such that

$$1 = \int_{\Omega} f dV \tag{A.23}$$

Hence, $f = 1/Vol(\Omega)$.

Example A.1: Uniform Density / Distribution on a Real Line Interval

Let a and b be constants with $a < b$. The uniform density on the interval (a,b) is the density f defined by

$$f(x) = \begin{cases} (b-a)^{-1} & \text{for } a < x < b, \\ 0 & \text{elsewhere} \end{cases} \tag{A.24}$$

The distribution function corresponding to (A.24) is given by

$$F(x) = \begin{cases} 0 & x < a, \\ (x-a)/(b-a), & a \leq x \leq b, \\ 1, & x > b. \end{cases} \tag{A.25}$$

Definition A.24: Binomial Density

Let $0 < p < 1$. Then, the real valued function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n, \\ 0, & \text{elsewhere} \end{cases} \tag{A.26}$$

is called the binomial density with parameters n and p.

Definition A.25: Geometric Density

Let $0 < p < 1$. Then the real valued function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} p(1-p)^x, & x = 0, 1, 2, ..., \\ 0, & \text{elsewhere} \end{cases} \tag{A.27}$$

is a discrete density function called the geometric density with parameter p.

Definition A.26: Poisson Density

Let $0 < p < 1$ and let λ be a positive number. Then, the real valued function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ..., \\ 0, & \text{elsewhere.} \end{cases} \tag{A.28}$$

is called the Poisson density with parameter λ.

Proposition A.4: Binomial Theorem

Let $0 < p < 1$ and $x < n \in \mathbb{Z}$. Then, we have that

$$1 = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{A.29}$$

Which follows from the binomial theorem

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} \tag{A.30}$$

Proposition A.5

Let ϕ be a differentiable strictly increasing or strictly decreasing function on an interval I, and let $\phi(I)$ denote the range of ϕ and ϕ^{-1} the inverse function to ϕ. Let X be a continuous random variable having density f such that $f(x) = 0$ for $x \notin I$. Then $Y = \phi(X)$ has density g given by $g(y) = 0$ for $y \notin \phi(I)$ and

$$g(y) = f\left(\phi^{-1}(y)\right)\left|\frac{d}{dy}\phi^{-1}(y)\right|, \qquad y \in \phi(I) \tag{A.31}$$

It is a bit more suggestive to wrtie this in the following form:

$$g(y) = f(x)\left|\frac{dx}{dy}\right|, \qquad y \in \phi(I) \qquad \text{and} \qquad x = \phi^{-1}(y) \tag{A.32}$$

Definition A.27: Cauchy Density

The following function f, is a density known as the Cauchy Density.

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty \tag{A.33}$$

The corresponding distribution function is given by

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x), \qquad -\infty < x < \infty \tag{A.34}$$

Definition A.28: Symmetric Density

A density function f is called symmetric *if $f(-x) = f(x)$ for all x. The Cauchy density and the uniform density on $(-a, a)$ are both symmetric.*

Proposition A.6

Let X be a random variable that has a density. Then f has a symmetric density if and only if X is a symmetric random variable.

Definition A.29: Standard Normal Density

The following density, ϕ

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty \tag{A.35}$$

The standard normal density is clearly symmetric.

The normal density with mean μ and variance σ^2 is often denoted by $n(\mu, \sigma^2)$ or $n(y; \mu, \sigma^2)$, $-\infty < y < \infty$. Thus,

$$n(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, \qquad -\infty < y < \infty \tag{A.36}$$

Definition A.30: Exponential Density

The exponential density with parameter λ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{A.37}$$

The corresponding distribution function is

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{A.38}$$

Proposition A.7

Let X be a random variable such that the following holds:

$$P(X > a + b) = P(X > a)P(X > b), \qquad a \geq 0 \qquad \text{and} \qquad b \geq 0 \tag{A.39}$$

Then either $P(X > 0) = 0$ or X is exponentially distributed.

Proposition A.8: Sum of Random Variables

Let X, Y be continuous random variables with densities f_X and f_Y respectively. Then, the random variable $Z = X + Y$ has density f_Z, given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-k) f_Y(k) \; dk \tag{A.40}$$

A.4 Expectations

Notation: Let $\mathbf{X}$ be a discrete r-dimensional random vector having possible values $\mathbf{x}_1, \mathbf{x}_2, \ldots$ and density f, and let ϕ be an real-valued function defined on $\mathbb{R}^r$. Then $\sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x})$ is defined as

$$\sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x}) = \sum_{j} \phi(\mathbf{x}_j) f(\mathbf{x}_j) \tag{A.41}$$

Definition A.31: Expectation Value

Let X be any discrete random variable that assumes a finite number of values $x_1, ..., x_r$. Then the expected value of X, denoted by EX, $E[X]$ or μ, is the number

$$E[X] = \sum_{i=1}^{r} x_i f(x_i) \tag{A.42}$$

The expected value $E[X]$ is also called the mean of X.

Definition A.32: Finite / Undefined Expectation

Let X be a discrete random variable having density f. If $\sum_j |x_j| f(x_j) < \infty$, then we say that X has finite expectation and we define its expectation by (A.42). On the other hand if $\sum_j |x_j| f(x_j) = \infty$, then we say X does not have finite expectation and $E[X]$ is undefined.

Proposition A.9

Let $\mathbf{X}$ be a discrete random vector having density f, and let ϕ be a real-valued function defined on $\mathbb{R}^r$. Then the random variable $Z = \phi(\mathbf{X})$ has finite expectation if and only if

$$\sum_{\mathbf{x}} |\phi(\mathbf{x})| f(\mathbf{x}) < \infty \tag{A.43}$$

and, when (A.43) holds,

$$E[Z] = \sum_{\mathbf{x}} \phi(\mathbf{x}) f(\mathbf{x}) \tag{A.44}$$

Proposition A.10: Properties of Expectation Operator

Let X and Y be two random variables having finite expectation.

(i) If c is a constant and $P(X = c) = 1$, then $E[X] = c$.
(ii) ***Linearity:***
a) If c is a constant, then cX has finite expectation and $E[cX] = cE[X]$.
b) $X + Y$ has finite expectation and[a]

$$E[X + Y] = E[X] + E[Y] \tag{A.46}$$

(iii) Suppose that $P(X \geq Y) = 1$. Then $E[X] \geq E[Y]$; moreover, $E[X] = E[Y]$ if and only if $P(X = Y) = 1$.
(iv) $|E[X]| \leq E[|X|]$.

[a] More explicitly, we note that these are expectations w.r.t different densities:

$$E_{X+Y}[X + Y] = E_X[X] + E_Y[Y] \tag{A.45}$$

Proposition A.11

Let X be a random variable such that for some constant M, $P(|X| \leq M) = 1$. Then X has finite expectation and $|E[X]| \leq M$.

Proposition A.12

Let X and Y be two independent random variables having finite expectations. Then XY has finite expectation and

$$E[XY] = E[X]E[Y] \tag{A.47}$$

Proposition A.13

Let X be a non-negative integer-valued random variable. Then X has finite expectation if and only if the series $\sum_{x=1}^{\infty} P(X \geq x)$ converges. If the series does converge, then

$$E[X] = \sum_{x=1}^{\infty} P(X \geq x) \tag{A.48}$$

Definition A.33: Moments / Central Moments

Let X be a discrete random variable, and let $r \geq 0$ be an integer. We say that X has a moment of order r if X^r has finite expectation. In that case we define the rth of X as $E[X^r]$.

If X has a moment of order r then the rth moment of $X - \mu$, where μ is the mean of X, is called the central moment (or the rth about the mean) of X.

Proposition A.14

If the random variables X and Y have moments of order r, then $X + Y$ also has a moment of order r.

Definition A.34: Variance

Let X be a random variable having a finite second moment. Then the variance of X, denoted by Var[X] or V[X], is defined by

$$Var[X] = E[(X - E[X])^2] \tag{A.49}$$

Through expanding, this works out to the following:

$$Var[X] = E[X^2] - (E[X])^2 \tag{A.50}$$

Definition A.35: Standard Deviation

We often denote Var X by σ^2. The non-negative number $\sigma = \sqrt{VarX}$ is called the standard deviation of X or of f_X.

Definition A.36: Covariance

Let X and Y be two random variables each having finite second moment. We define a quantity called the covariance of X and Y written as Cov(X, Y). Thus we have the formula

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \tag{A.51}$$

We note that $X + Y$ has a finite second moment and finite variance. We therefore have an important formula:

$$Var[X + Y] = Var[X] + Var[Y] + 2\ Cov(X, Y) \tag{A.52}$$

Definition A.37: Correlation Coefficient

Let X and Y be two random variables having finite nonzero variances. One measure of the degree of dependence between the two random variables is the correlation coefficient $\rho(X, Y)$ defined by

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{(Var[X])(Var[Y])}} \tag{A.53}$$

These random variables are said to be uncorrelated if $\rho = 0$. We can automatically see that independent random variables are uncorrelated. However, it is possible for dependent random variables to be uncorrelated as well. We observe that the correlation coefficient ρ is always between -1 and 1, and that $\rho = 1$ if and only if $P(X = aY) = 1$ for some constant a.

Definition A.38: Cross-Correlation Matrix

Let $\mathbf{X}$, $\mathbf{Y}$ be random vectors. Then, we define the cross-correlation matrix by $E[\mathbf{X}\mathbf{Y}^T]$, where the matrix elements in the standard basis are given by $\left|E[\mathbf{X}\mathbf{Y}^T]\right|_{ij} = E[x_i y_j]$.

Theorem A.1: The Schwartz Inequality

Let X and Y have finite second moments. Then

$$[E[XY]]^2 \leq (E[X^2])(E[Y^2]) \tag{A.54}$$

Furthermore, equality holds in (A.54) if and only if either P(Y=0) = 1 or P(X = aY) = 1 for some constant a.

Proposition A.15: Chebyshev's Inequality

Let X be a random variable with mean μ and finite variance σ^2. Then for any real number $t > 0$

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{A.55}$$

Theorem A.2: Weak Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be independent random variables having acommon distribution with finite mean μ and set $S_n = X_1 + ... + X_n$. Then for any $\delta > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \delta\right) = 0 \tag{A.56}$$

A.5 Jointly Distributed Random Variables

Definition A.39: Joint Distribution Function

Let X and Y be two random variables defined on the same probability space. Their joint distribution function F is defined by

$$F(x,y) = P(X \leq x, Y \leq y), \qquad -\infty < x, y < \infty \tag{A.57}$$

Definition A.40: Marginal Distribution Functions

The one-dimensional distribution functions F_X and F_Y defined by

$$F_X(x) = P(X \leq x) \qquad \text{and} \qquad F_Y(y) = P(Y \leq y) \tag{A.58}$$

are called the marginal distribution functions of X and Y. They are related to the joint distribution function F by

$$F_X(x) = F(x, \infty) = \lim_{y \to \infty} F(x, y) \tag{A.59}$$

$$F_Y(y) = F(\infty, y) = \lim_{x \to \infty} F(x, y) \tag{A.60}$$

Definition A.41: Joint Density Function

If there is a nonnegative function f such that

$$F(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f(u, v) dv \right) du, \qquad -\infty < x, y < \infty, \tag{A.61}$$

then f is called a joint density function (with respect to integration) for the distribution function F or the pair of random variables X, Y.

$$P((X, Y) \in A) = \int_A \int f(x, y) \; dx \; dy \tag{A.62}$$

By letting A be the entire plane we obtain

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \; dx \; dy = 1 \tag{A.63}$$

Definition A.42: Marginal Density

Let F be the distribution function for a pair of random variables X, Y. Then, the marginal distribution F_X has marginal density f_X given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) \; dy \tag{A.64}$$

Similarly, F_Y has marginal density f_Y given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \; dx \tag{A.65}$$

We note that it satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(u) \; du \tag{A.66}$$

We can observe that

$$\frac{\partial^2}{\partial x \; \partial y} F(x, y) = f(x, y) \tag{A.67}$$

Definition A.43: Independent Random Variables

The variables X and Y are called independent random variables if whenever $a \leq b$ and $c \leq d$, then

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d) \tag{A.68}$$

By letting $a = c = -\infty$, $b = x$, and $d = y$, it follows that if X and Y are independent, then

$$F(x, y) = F_X(x)F_Y(y), -\infty < x, y < \infty \tag{A.69}$$

Proposition A.16

If X and Y are independent and A and B are unions of a finite or countably infinite number of intervals, then

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{A.70}$$

In other words, the events

$$\{\omega | X(\omega) \in A\} \qquad \text{and} \qquad \{\omega | X(\omega) \in B\} \tag{A.71}$$

are independent events.

Proposition A.17

Let X and Y be random variables having marginal densities f_X and f_Y. Then X and Y are independent if and only if the function f defined by

$$f(x, y) = f_X(x)f_Y(y), \qquad -\infty < x, y < \infty \tag{A.72}$$

is a joint density for X and Y.

Definition A.44: Bivariate Density Function

A two-dimensional (or bivariate)density function f is a non-negative function on $\mathbb{R}^2$ such that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1 \tag{A.73}$$

Definition A.45: Standard Bivariate Normal Density

The density given below by f is referred to as the standard bivariate normal density.

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}, \qquad -\infty < x, y < \infty \tag{A.74}$$

Proposition A.18

Let X and Y be random variables having joint density f. In many contexts, we will have a random variable Z defined in terms of X and Y and we wish to calculate the density of Z. Let $Z = \phi(X, Y)$, where ϕ is a real-valued function whose domains contains the range of X and Y. For fixed z the event $\{Z \leq z\}$ is equivalent to the event $\{(X, Y) \in A_z\}$ where A_z is the subset of $\mathbb{R}^2$ defined by

$$A_z = \{(x, y) | \phi(x, y) \leq z\} \tag{A.75}$$

Thus,

$$F_Z(z) = P(Z \leq z) \tag{A.76}$$
$$= P((X, Y) \in A_z) \tag{A.77}$$
$$= \int_{A_z}\int f(x, y)\ dx\ dy \tag{A.78}$$

If we can find a non-negative function g such that

$$\int_{A_z}\int f(x, y)\ dx\ dy = \int_{-\infty}^{z} g(v)\ dv, \qquad -\infty < z < \infty \tag{A.79}$$

then g is necessarily a density of Z.

Proposition A.19

Let X and Y be independent random variables having the respective normal densities $n(\mu_1, \sigma_1^2)$ *and* $n(\mu_2, \sigma_2^2)$*. Then X + Y has the normal density*

$$n(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{A.80}$$

Definition A.46: Conditional Density (Discrete)

Let X and Y be discrete random variables having joint density f. If x is a possible value of X, then

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f(x, y)}{f_X(x)} \tag{A.81}$$

The function $f_{Y|X}$ *defined by*

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & f_X(x) \neq 0 \\ 0, & f_X(x) = 0 \end{cases} \tag{A.82}$$

is called the conditional density of Y given x.

Definition A.47: Conditional Density (Continuous)

Let X and Y be continuous random variables having joint density f. The conditional density $f_{Y|X}$ *is defined by*

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & 0 < f_X(x) < \infty, \\ 0 & \textit{elsewhere.} \end{cases} \tag{A.83}$$

If f_X *is continuous and* $f_X(x) \neq 0$*, we have*

$$P(a \leq Y \leq b|X = x) = \frac{\int_a^b f(x, y)\ dy}{f_X(x)} \tag{A.84}$$

Proposition A.20: Bayes Rule

Let X and Y be random variables with marginal densities f_X and f_Y respectively and conditional densities $f_{X|Y}$ and $f_{Y|X}$. We have the continuous analog to Bayes' rule given below:

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x)\ dx} \tag{A.85}$$

Definition A.48: Joint Distribution Function (Multivariate)

Let $X_1, ..., X_n$ be n random variables defined on a common probability space. Their joint distribution function F is defined by

$$F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n), \qquad -\infty < x_1, ..., x_n < \infty \tag{A.86}$$

Definition A.49: Marginal Distribution Function (Multivariate)

The marginal distribution functions $F_{X_m}, m = 1, ..., n$ are defined by

$$F_{X_m}(x_m) = P(X_m \leq x_m), \qquad -\infty < x_m < \infty \tag{A.87}$$

The value of $F_{X_m}(x_m)$ can be obtained from F by letting $x_1, ..., x_{m-1}, x_{m+1}, ..., x_n$ all approach $+\infty$.

Definition A.50: Joint Density Function (Multivariate)

A non-negative function f is called a joint density function (with respect to integration) for the joint distribution function F, or for the random variables $X_1, ..., X_n$ if

$$F(x_1, ..., x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, ..., u_n)\ du_1 \cdots du_n, \qquad -\infty < x_1, ..., x_n < \infty \tag{A.88}$$

We also note that

$$f(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, ..., x_n) \tag{A.89}$$

is valid at the continuity points of F.

Definition A.51: Marginal Density Function (Multivariate)

The random variable X_m has the marginal density f_{X_m} obtained by integrating f over the remaining $n-1$ variables. For example,

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n)\ dx_1\ dx_3 \cdots dx_n \tag{A.90}$$

Definition A.52: Independent Random Variables (Multivariate)

In general, the random variables $X_1, ..., X_n$ are called independent whenever $a_m \leq b_m$ for $m = 1, ..., n$, then

$$P(a_1 < X_1 \leq b_1, ..., a_n < X_n \leq b_n) = P(a_1 < X_1 \leq b_1) \cdots P(a_n < X_n \leq b_n) \tag{A.91}$$

Proposition A.21

A necessary and sufficient condition for independence is that

$$F(x_1, ..., x_n) = F_{x_1}(x_1) \cdots F_{X_n}(x_n), \qquad -\infty < x_1, ..., x_n < \infty \tag{A.92}$$

If F has a density f, then $X_1, ..., X_n$ are independent if and only if f can be chosen so that

$$f(x_1, ..., x_n) = f_{x_1}(x_1) \cdots f_{X_n}(x_n), \qquad -\infty < x_1, ..., x_n < \infty \tag{A.93}$$

If $X_1, ..., X_n$ are random variables whose joint density is given by (A.93) then $X_1, ..., X_n$ are independent and X_m has the marginal density f_m.

Proposition A.22

Let $X_1, ..., X_n$ be independent random variables. Let Y be a random variable defined in terms of $X_1, ..., X_m$ and let Z be a random variable defined in terms $X_{m+1}, ..., X_n$ (where $1 < m < n$). Then Y and Z are independent.

Definition A.53: Conditional Density (Multivariate)

If $X_1, ..., X_n$ has a joint density f, then any subcollection of these random variables has a joint density which can be found by integrating over the remaining variables. For example, if $1 \leq m < n$,

$$f_{X_1,...,X_m}(x_1, ..., x_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n) dx_{m+1} \cdots dx_n \tag{A.94}$$

The conditional density of a subcollection of $X_1, ..., X_n$ given the remaining variables can also be defined in an obvious manner. Thus the conditional density of $X_{m+1}, ..., X_n$ given $X_1, ..., X_m$ is defined by

$$f_{X_{m+1},...,X_n|X_1,...,X_m}(x_{m+1}, ..., x_n|x_1, ..., x_m) = \frac{f(x_1, ..., x_n)}{f_{X_1,...,X_m}(x_1, ..., x_m)} \tag{A.95}$$

Where f is the joint density of $X_1, ..., X_n$.

Definition A.54: Order Statistics

Let $U_1, ...U_n$ be independent continuous random variables, each having distribution F and density function f. Let $X_1, ..., X_n$ be random variables obtained by letting $X_1(\omega), ...X_n(\omega)$ be the set $U_1(\omega), ..., U_n(\omega)$ permuted so as to be in increasing order. In particular, we define X_1 and X_n to be the functions

$$X_1(\omega) = min\{U_1(\omega), ..., U_n(\omega)\} \tag{A.96}$$

and

$$X_n(\omega) = max\{U_1(\omega), ..., U_n(\omega)\} \tag{A.97}$$

The random variable X_k is called the k^{th} order statistic. Another related variable of interest is the range R, defined by

$$R(\omega) = X_n(\omega) - X_1(\omega) \tag{A.98}$$
$$= max\{U_1(\omega), ..., U_n(\omega)\} - min\{U_1(\omega), ..., U_n(\omega)\} \tag{A.99}$$

Proposition A.23: Distributions for Order Statistics

Let $X_1, X_2, ..., X_n$ be identically distributed and independent random variables. Let their common CDF by denoted by F. We define $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ as the vector of order statistics of $X_1, X_2, ..., X_n$. Then, the distribution for $X_{(k)}$ in a sample of size n is given by

$$F_{(k,n)}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{k} [F(x)]^j [1 - F(x)]^{n-j} \tag{A.100}$$

Proof. We will break the event $(X_{(k)} \leq x)$ into disjoint sub-events given by

$$(X_{(k)} \leq x) = (X_{(n)} \leq x) \cup (X_{(n)} > x, X_{(n-1)} \leq x) \cup ... \cup (X_{(n)} > x, ..., X_{(k+1)} > x, X_{(k)} \leq x). \tag{A.101}$$

Recall from the property of probability measures that if $A_i' s$ are all mutually disjoint sets, then $P(\cup_{i=1}^{l} A_i) = \sum_{i=1}^{l} P(A_i)$. Hence, it amounts to identify the probability of the events contained within each of these terms. Consider the event $(X_{(n)} > x, ..., X_{(j+1)} > x, X_{(j)} \leq x)$. This event tells us that the first j ordered variables have a value less than x while the rest have a value lying above x. Since the CDF, F(x) to each one tells us the probability of them occupying a value less than x, one can view this through the lens of fail / successes among n independent variables. In essence, we use the multiplicative property of independent events and combinatorics to establish the number of combinations one can arrange such an ordering of variables. We therefore have

$$P(X_{(n)} > x, ..., X_{(j+1)} > x, X_{(j)} \leq x) = \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}. \tag{A.102}$$

Hence, it therefore follows that

$$F_{(k,n)}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j} \tag{A.103}$$

□

Theorem A.3: Change of Variables

Let $X_1, ..., X_n$ be continuous random variables having joint density f and let random variables $Y_1, ..., Y_n$ be defined by

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j, \qquad i = 1, ..., n, \tag{A.104}$$

where the matrix $A = [a_{ij}]$ has nonzero determinant det A. Then $Y_1, ..., Y_n$ have joint density $f_{Y_1,...,Y_n}$ given by

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = \frac{1}{|det\ A|} f(x_1, ..., x_n), \tag{A.105}$$

where the x's are defined in terms of y's as the unique solution to the equations $y_i = \sum_{j=1}^{n} a_{ij} x_j$.

A.6 Expectations and the Central Limit Theorem

Definition A.55: Expectation (Continuous)

Let X be a continuous random variable having density f. We say that X has finite expectation if

$$\int_{-\infty}^{\infty} |x| f(x)\ dx < \infty, \tag{A.106}$$

and in that case we define its expectation by

$$E[X] = \int_{-\infty}^{\infty} x f(x)\ dx \tag{A.107}$$

Theorem A.4

Let $X_1, ..., X_n$ be continuous random variables having joint density f and let Z be a random variable defined in terms of $X_1, ..., X_n$ be $Z = \phi(X_1, ..., X_n)$. Then Z has finite expectation if and only if

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\phi(x_1, ..., x_n)| f(x_1, ..., x_n)\ dx_1 \cdots dx_n < \infty \tag{A.108}$$

in which case

$$E[Z] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(x_1, ..., x_n) f(x_1, ..., x_n)\ dx_1 \cdots dx_n < \infty \tag{A.109}$$

Definition A.56: Moments (Continuous)

Let X be a continuous random variable having density f and mean μ. If X has finite mth moment, then we have

$$E[X^m] = \int_{-\infty}^{\infty} x^m f(x)\ dx \tag{A.110}$$

If X has finite second moment, its variance σ^2 is given by

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx \tag{A.111}$$

Definition A.57: Conditional Expectation

Let X and Y be continuous random variables having joint density f and suppose that Y has finite expectation. Recall that we defined the conditional density of Y given $X = x$ by

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & 0 < f_X(x) < \infty, \\ 0 & \textit{elsewhere.} \end{cases} \tag{A.112}$$

For each x such that $0 < f_X(x) < \infty$ the function $f_{Y|X}(y|x)$, $-\infty < y < \infty$, is a density function with respect to ***Def*** *A.22. Thus we can talk about various moments of this density. Its mean is called the conditional* *expectation of Y given $X = x$ and is denoted by $E[Y|X = x]$ or $E[Y|x]$. Thus*

$$E[Y|X = x] = \int_{-\infty}^{\infty} yf(y|x)\ dy \tag{A.113}$$

$$= \frac{\int_{-\infty}^{\infty} yf(x,y)\ dy}{f_X(x)} \tag{A.114}$$

when $0 < f_X(x) < \infty$. We define $E[Y|X = x] = 0$ elsewhere.

Proposition A.24: Properties of Conditional Expectation

Let X, Y, Z be random variables and $a, b \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Assuming all the following expectations exist, we have that

(i) $E[a|Y] = a$
(ii) $E[aX + bY|Z] = aE[X|Y] + bE[Y|Z]$
(iii) $E[X|Y] \geq 0$ *if* $X \geq 0$.
(iv) $E[X|Y] = E[X]$ *if* X *and* Y *are independent.*
(v) $E[E[X|Y]] = E[X]$
(vi) $E[Xg(Y)|Y] = g(Y)E[X|Y]$. *In particular,* $E[g(Y)|Y] = g(Y)$.
(vii) $E[X|Y, g(Y)] = E[X|Y]$
(viii) $E[E[X|Y, Z]|Y] = E[X|Y]$

Definition A.58: Regression Function

In statistics, the function m defined by $m(x) = E[Y|X = x]$ is called the regression function of Y on X.

Lemma A.1

Let X be a random variable with density f_X. Then, $\frac{X-\alpha}{\beta}$ is a random variable with density

$$f_{(X-\alpha)/\beta}(z) = \beta f_X(\beta z + \alpha) \tag{A.115}$$

Theorem A.5: Central Limit Theorem

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables having mean μ and finite nonzero variance σ^2. Set $S_n = X_1 + \cdots + X_n$. Then

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \qquad -\infty < x < \infty \tag{A.116}$$

Where we recall that Φ is the CDF for the normal density:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\ du \tag{A.117}$$

Proof. Let $X_1, X_2, ...$ be independent, identically distributed random variables having mean μ and finite nonzero variance σ^2. Let their density be denoted by f. We define the random variable $S_n = X_n + S_{n-1}$, noting that its density is given by

$$f_{S_n}(x_n) = \int_{-\infty}^{\infty} \cdot \cdot \int_{-\infty}^{\infty} \mathrm{d}x_1 \ \mathrm{d}x_2 \cdot \cdot \mathrm{d}x_{n-1} \left(\prod_{i=1}^{n-1} f(x_{i+1} - x_i) \right) f(x_1) \tag{A.118}$$

We define a new random variable $G_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Its density is given by

$$f_{G_n}(x_n) = \sigma\sqrt{n} \int_{-\infty}^{\infty} \cdot \cdot \int_{-\infty}^{\infty} \mathrm{d}x_1 \ \mathrm{d}x_2 \cdot \cdot \mathrm{d}x_{n-1} \ f(\sigma\sqrt{n}x_n + n\mu - x_{n-1}) \left(\prod_{i=1}^{n-2} f(x_{i+1} - x_i) \right) f(x_1) \tag{A.119}$$

□

B Linear Algebra

Proposition B.1

Let X be a $n \times p$ matrix. Then $(X^TX)^T = X^TX$ and if X^TX is invertible, we then have $((X^TX)^{-1})^T = (X^TX)^{-1}$.

Proof. The first statement is trivial. For the second statement, we observe that

$$\begin{aligned} (X^TX)(X^TX)^{-1} &= 1 \\ ((X^TX)(X^TX)^{-1})^T &= 1 \\ ((X^TX)^{-1})^T(X^TX) &= 1 \\ ((X^TX)^{-1})^T &= (X^TX)^{-1} \end{aligned} \tag{B.1}$$

□

Proposition B.2: Recasting Weighted Sum Into a Matrix Equation

Let $x^{(i)} \in \mathcal{M}_{1\times n_x}$ be a vector and let $v_i \in \mathbb{R}$ where $1 \leq i \leq m$. Then, we define the $n_x \times m$ matrix X such that $x^{(i)}$ are stacked beside each other in columns[a]:

$$X = \begin{bmatrix} | & | & \dots & | \\ | & | & \dots & | \\ x^{(1)} & x^{(2)} & \dots & x^{(m)} \\ | & | & \dots & | \\ | & | & \dots & | \end{bmatrix} \tag{B.2}$$

Hence, the matrix elements are given by $X_{ij} = x_i^{(j)}$ with $x_i^{(j)}$ indicating the i^{th} entry of the vector $x^{(j)}$. Similarly, we define a $1 \times m$ vector V by

$$V = \begin{bmatrix} v_1 & v_2 & \dots & v_m \end{bmatrix} \tag{B.3}$$

We now establish the following identity:

$$B = \sum_{i=1}^{m} v_i \, x^{(i)} = XV^T \tag{B.4}$$

Proof. We consider both these objects equivalent if their underlying elements are the same. We observe that B is a $m \times 1$ vector. Hence, we have that

$$B_j = \sum_{i=1}^{m} v_i \; x_j^{(i)}. \tag{B.5}$$

Similarly, we observe that XV^T is a $m \times 1$ vector. We note that its elements are given by

$$\Big[XV^T\Big]_j = \sum_{i=1}^{m} X_{ji} V_i = \sum_{i=1}^{m} x_j^{(i)} v_i = B_j. \tag{B.6}$$

We therefore have the identity:

$$\sum_{i=1}^{m} v_i \; x^{(i)} = XV^T. \tag{B.7}$$

□

[a] If $x^{(i)}$ correspond to the input elements of a training set, then X is the training matrix, defined by Def **??**.

Definition B.1: Positive / Negative Definite Matrix

Let M be an $n \times n$ symmetric real matrix. Then, we say that M is positive definite iff the scalar

$$x^T M x > 0 \;\; \forall x \in \mathbb{R}^n \setminus \mathbf{0}. \tag{B.8}$$

Similarly, we say that M is negative definite iff the scalar

$$x^T M x < 0 \;\; \forall x \in \mathbb{R}^n \setminus \mathbf{0} \tag{B.9}$$