Contents

1 Probability theory and stochastic processes 1

1.1 Probability space 1
1.2 Random variables and their expectations 2
1.3 Variance, covariance, and correlation 4
1.4 Probability distribution 6
1.5 Examples of discrete probability distributions 9
1.6 Examples of continuous probability distributions 12
1.7 Independence 15
1.8 Conditional Probability and Expectation 17
1.9 Stochastic processes 20
1.10 Summary 25
1.11 Questions 27

2 Claim size estimation in insurance and reinsurance 28

2.1 Basic principles of insurance risk modelling 28
2.2 Method of moments 31
2.3 Method of maximum likelihood 34
2.4 Method of percentiles 37
2.5 Reinsurance 39
2.6 Claim size estimation with excess of loss reinsurance 41
2.7 Summary 45
2.8 Questions 47

3 Estimation of aggregate claim distribution 48

3.1 The collective risk model 48
3.2 The compound Poisson distribution 51
3.3 The compound binomial distribution 54
3.4 The compound negative binomial distribution 56
3.5 Aggregate claim distribution under reinsurance 57
3.6 The individual risk model 60
3.7 Aggregate claim estimation under uncertainty in parameters . 62
3.8 Summary 66
3.9 Questions 68

4 Tails and dependence analysis of claims distributions 69

4.1 How likely very large claims to occur? 69
4.2 The distribution of large claims 72
4.3 The distribution of maximal claim 73
4.4 Dependence, correlation, and concordance 75

4.5 Joint distributions and copulas 77
4.6 Dependence of distribution tails 81
4.7 Summary 83
4.8 Questions 85

5 Markov Chains 85
5.1 The Markov property 86
5.2 Definition of Markov Chains 87
5.3 The Chapman-Kolmogorov equations 89
5.4 Time dependency of Markov chains 90
5.5 Further applications 92
5.5.1 The simple (unrestricted) random walk 92
5.5.2 The restricted random walk 94
5.5.3 The modified NCD model 96
5.5.4 A model of accident proneness 98
5.5.5 A model for credit rating dynamics 98
5.5.6 General principles of modelling using Markov chains . . 99
5.6 Stationary distributions 101
5.7 The long-term behaviour of Markov chains 104
5.8 Summary 105
5.9 Questions 106

6 Markov Jump Processes 108
6.1 Poisson process 108
6.1.1 Interarrival times 110
6.1.2 Compound Poisson process 112
6.2 The time-inhomogeneous Markov jump process 113
6.3 Transition rates 114
6.4 Time-homogeneous Markov jump processes 117
6.5 Applications 119
6.5.1 Survival model 119
6.5.2 Sickness-death model 120
6.5.3 Marriage model 122
6.6 Summary 124
6.7 Questions 125

7 Machine Learning 126
7.1 A motivating example 126
7.2 The problems machine learning can solve 129
7.3 Models, methods, and techniques 132
7.4 Probabilistic analysis 135

7.5 Stages of analysis in Machine Learning 137
7.6 Summary 141
7.7 Questions 143

8 Solutions of end-of-chapter questions 144
8.1 Chapter 1 solutions 144
8.2 Chapter 2 solutions 148
8.3 Chapter 3 solutions 151
8.4 Chapter 4 solutions 154
8.5 Chapter 5 solutions 159
8.6 Chapter 6 solutions 164
8.7 Chapter 7 solutions 168

1.10 Summary

A random variable is a measurable function $\xi : \Omega \to \mathbb{R}$ from probability space Ω to real line.

For a random variable ξ, its cumulative distribution function (cdf) is $F_\xi(x) := P(\xi \leq x)$. If $F_\xi(x)$ can be represented as

$$F_\xi(x) = \int_{-\infty}^{x} \rho_\xi(z)\, dz,$$

for some non-negative function ρ_ξ, the latter is called the probability density function (pdf) of ξ.

The expectation of a random variable ξ is defined as $E[\xi] \equiv \int_\Omega \xi\, dP$. It can be calculated as

$$E[\xi] = \int_{-\infty}^{\infty} x\, dF_\xi(x) = \int_{-\infty}^{\infty} x \rho_\xi(x)\, dx.$$

The variance of a random variable ξ is defined by $\mathrm{Var}(\xi) := E[(\xi - E[\xi])^2] \equiv E[\xi^2] - E[\xi]^2$. For two random variables ξ, η, their covariance is defined by $\mathrm{Cov}(\xi, \eta) := E\big[(\xi - E[\xi])(\eta - E[\eta])\big] \equiv E[\xi\eta] - E[\xi]E[\eta]$.

Two events A and B are called independent if $P(A \cap B) = P(A)P(B)$. Random variables ξ and η are called *independent* if the events $A = \{\omega \in \Omega : a < \xi(\omega) < b\}$ and $B = \{\omega \in \Omega : c < \eta(\omega) < d\}$ are independent for all real numbers a, b, c, d. If ξ and η are independent, $\mathrm{Cov}(\xi, \eta) = 0$, but the converse is not always true.

The conditional probability of A given B, denoted $P[A|B]$, can be evaluated as $P[A|B] = \frac{P[A \cap B]}{P[B]}$.

A stochastic process is a family of random variables indexed in time, $\{X_t : t \in \mathcal{T}\}$. The time set $\mathcal{T}$ can be discrete or continuous, as can the state space $\mathcal{S}$ in which the variables take their values.

Stochastic process can be roughly classified into the following groups:

- Discrete $\mathcal{S}$ and discrete $\mathcal{T}$;
- Continuous $\mathcal{S}$ and discrete $\mathcal{T}$;
- Discrete $\mathcal{S}$ and continuous $\mathcal{T}$;
- Continuous $\mathcal{S}$ and continuous $\mathcal{T}$; or

- Mixed processes.

A stochastic process $\{X_t : t \in \mathcal{T}\}$ is said to be *stationary*, if for all integers n and all $t, t_1, t_2, \ldots, t_n$ in $\mathcal{T}$ the joint distributions of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ and $X_{t+t_1}, X_{t+t_2}, \ldots, X_{t+t_n}$ coincides. It is called *weakly stationary*, if the mean of the process, $m(t) = E[X_t]$, is constant, and the covariance of the process, $C(s,t) = E[(X_s - m(s))(X_t - m(t))]$, depends only on the time difference $t - s$.

An *increment* of the stochastic process $\{X_t : t \in \mathcal{T}\}$ is the quantity $X_{t+u} - X_t$, $u > 0$. A stochastic process has stationary (or time-homogeneous) increments, if for every $u > 0$ the increment $Z_t = X_{t+u} - X_t$ is a stationary process; a process $\{X_t : t \in \mathcal{T}\}$ is said to have independent increments if for any $a, b, c, d \in \mathcal{T}$ such that $a < b < c < d$, random variables $X_a - X_b$ and $X_c - X_d$ are independent.

2.7 Summary

In this chapter we focus on modelling the claim size distribution. We assume that claim distribution belongs to certain family, but with unknown parameters. The company estimate unknown parameters $a_1, a_2, \ldots, a_r$ to fit the data of past claims $x_1, x_2, \ldots, x_n$ as good as possible.

The method of moments suggests to select parameters in such a way that first r moments estimated from the data match the first r moments $f_j(a_1, a_2, \ldots, a_r)$, $j = 1, 2, \ldots, r$ estimated from the formulas for the distribution, that is,

$$m_j = f_j(a_1, a_2, \ldots, a_r), \quad j = 1, 2, \ldots, r$$

where $m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$, $j = 1, 2, \ldots, r$.

Method of maximum likelihood suggests to select vector of parameters $theta = (a_1, a_2, \ldots, a_r)$ to maximize (the logarithm of) the likelihood function

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log[f(x_i \,|\, \theta)].$$

The optimal $\hat{\theta} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_r)$ can be found from the system of equations

$$\frac{d}{da_i} l(\hat{\theta}) = 0, \quad i = 1, 2, \ldots, r.$$

Method of percentiles suggests to find vector of parameters $\lambda = (a_1, a_2, \ldots, a_r)$ from the system of equations

$$\alpha_i = F(q_i, \lambda), \quad i = 1, 2, \ldots, r,$$

where F is a cdf which depends on parameters, $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_r < 1$ are some pre-specified numbers, and q_i is the estimate of the percentile at level α_i based on data $x_1, x_2, \ldots, x_n$. To find it, we first find the smallest integer j greater than $n\alpha_i$, then sort the sequence $x_1, x_2, \ldots x_n$ is non-decreasing order, and then the j-th smallest number is q_i.

To protect itself from large claims, an insurance company may in turn take out an insurance policy in another company, called reinsurer. We consider reinsurance contracts of two very simple types: proportional reinsurance and individual excess of loss reinsurance. In proportional reinsurance, if claim amount is X, then insurer pays αX and reinsurer pays $(1 - \alpha)X$, where α is the parameter known as the retention level. With excess of loss reinsurance contract, if the claim for amount X arrives then the insurer will pay

$$Y = \begin{cases} X & \text{if } X \leq M \\ M & \text{if } X > M \end{cases}$$

and the reinsurer pays $Z = X - Y$.

3.8 Summary

Let N be a (random) number of claims during some fixed period of time. If the sizes of claims are $X_1, X_2, \dots, X_N$, then the total cost to cover all claims is

$$S = X_1 + X_2 + \cdots + X_N,$$

and $S = 0$ if $N = 0$. If all X_i are independent, identically distributed, and also independent of N, then we say that S has *compound distribution.*

The cdf of S is given by

$$G(x) = P(S \le x) = \sum_{n=0}^{\infty} P(N = n) \cdot F^{n*}(x),$$

where $F(x)$ is the cdf of the individual claim and $F^{n*}(x)$ its n-fold convolution.

We denote μ_X, μ_N, μ_S the means of random variables X_i, N, S and by σ_X^2, σ_N^2, σ_S^2 the corresponding variances. Then

$$\mu_S = \mu_N \cdot \mu_X, \quad \text{and} \quad \sigma_S^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2.$$

For example,

- If N follow a Poisson distribution with parameter λ, then S is called compound Poisson and

$$\mu_S = \lambda \mu_X, \quad \text{and} \quad \sigma_S^2 = \lambda E[X^2].$$

- If N follow a binomial distribution with parameter n and p, then S is called compound binomial and

$$\mu_S = np\mu_X. \quad \text{and} \quad \sigma_S^2 = npE[X^2] - np^2\mu_X^2.$$

- If N follow a negative binomial distribution with parameters k and p, then S is called compound binomial and

$$\mu_S = \frac{k(1-p)}{p}\mu_X \quad \text{and} \quad \sigma_S^2 = \frac{k(1-p)}{p}E[X^2] + \frac{k(1-p)^2}{p^2}\mu_X^2.$$

Under proportional reinsurance with retention level α, the i-th individual claim amount for the insurer is $Y_i = \alpha X_i$. Under the excess of loss reinsurance with retention level M, it is $Y_i = \min(X_i, M)$. In both cases, the aggregate claim for the insurer is

$$S_I = Y_1 + Y_2 + \cdots + Y_N,$$

and all the formulas above work with Y_i in place of X_i.

For the reinsurer, the individual claim amount is $Z_i = X_i - Y_i = (1-\alpha)X_i$ under proportional reinsurance and $Z_i = X_i - Y_i = \max(0, X_i - M)$ under the excess of loss reinsurance. In both cases, the aggregate claim for the reinsurer is

$$S_R = Z_1 + Z_2 + \cdots + Z_N,$$

and all the formulas above work with Z_i in place of X_i.

In the individual risk model, the aggregate claim is

$$S = Y_1 + Y_2 + \cdots + Y_n,$$

where $Y_j = N_j X_j$, $N_j \sim Bin(1, q_j)$ is the number of claims from j-th policy, and X_j is the individual claim amount with mean m_j and variance σ_j^2. Then

$$E[S] = \sum_{j=1}^{n} q_j \mu_j, \quad \text{and} \quad Var[S] = \sum_{j=1}^{n} (q_j \sigma_j^2 + q_j(1-q_j)\mu_j^2).$$

4.7 Summary

The "heavier" is the tail of claim distribution, the more likely very large claims to occur. We can "quantify" how heavy are the tails by analysing the following:

- Existence of moments $M_k = \int_0^\infty x^k f(x)dx$, where $f(x)$ is a density of a non-negative r.v.
- Limiting density ratio: $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)}$, where $f(x)$ and $g(x)$ are two densities.
- Hazard rate $h(x) = \frac{f(x)}{1-F(x)}$, where $f(x)$ is density and $F(x)$ is cdf.
- Mean residual life $e(x) = \frac{\int_x^\infty (y-x)f(y)dy}{\int_x^\infty f(y)dy}$.

Let X be a random variable with cumulative distribution function F, $u \in \mathbb{R}$ be a threshold, and let

$$F_u(x) = P(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

Then

$$\lim_{u\to\infty} F_u(x) = G(x),$$

where $G(x)$ belongs to a two-parameter family of Generalised Pareto distributions.

Let

$$M_n = \max\{X_1, X_2, \dots, X_n\}.$$

where $X_1, X_2, \dots, X_n$ are i.i.d. random variables. The extreme value theorem states that if $a_1, a_2, \dots, a_n, \dots$ and $b_1, b_2, \dots, b_n, \dots$ are two sequence of real numbers such that the limit

$$F(x) = \lim_{n\to\infty} P\left(\frac{M_n - a_n}{b_n} \leq x\right)$$

exists and depends only on x, then $F(x)$ must follow the Generalized Extreme Value (GEV) distribution. It has 3 parameters: location parameter α, scale parameter $\beta > 0$, and shape parameter γ.

If $\gamma < 0$, $\gamma = 0$, and $\gamma > 0$, the GEV distribution reduces to the Weibull distribution, the Gumbel distribution, and the Fretchet distribution, respectively.

For 2 random variables X and Y, the "concordance" measures to what extend we have direct dependence of the form "the higher X the higher

Y" (this corresponds to positive concordance), and to what extend we have the opposite dependence "the higher X the lower Y" (corresponding to the negative concordance). If X and Y are independent, the concordance is 0. There are several ways to measure concoradance, one of them is the Kendall coefficient

$$\tau = \frac{C - D}{C + D},$$

where C and D are the numbers on concordant and discordant pairs of scenarios, respectively.

The joint cdf F of n random variables can be written in the form

$$F(x_1, x_2, \dots, x_n) = C(F_1(x_1), F_2(x_2), \dots, F_n(x_n)),$$

where $F_i(x_i)$ are cdfs of individual random variables (known as marginal distributions), and function C is called a copula and represents the way random variables depend on each other.

The Archimedean family is the family of copulas in the form:

$$C(x_1, x_2, \dots, x_n) = \psi^{-1}(\psi(x_1) + \psi(x_2) + \dots + \psi(x_n)),$$

where $\psi : (0, 1] \to [0, \infty)$ is a continuous, strictly decreasing, convex function with $\psi(1) = 0$.

For example,

$$C(x_1, x_2, \dots, x_n) = x_1 \cdot x_2 \cdot \dots \cdot x_n$$

is the independence copula, which implies that all n random variables are jointly independent, while

$$C(x_1, x_2, \dots, x_n) = \min(x_1, x_2, \dots, x_n)$$

is the co-monotonic copula, which corresponds to the case when all variables are directly dependent (the higher is one variable, the higher are all). For $n = 2$,

$$C(x_1, x_2) = \max(x_1 + x_2 - 1, 0)$$

is called counter-monotonic copula, which represents the inverse dependence of the form "the higher one variable, the lower another one".

The coefficients of lower and upper tail dependence of 2 random variables are

$$\lambda_L = \lim_{u \to 0+} \frac{C(u, u)}{u}, \quad \text{and} \quad \lambda_U = \lim_{u \to 0+} \frac{\bar{C}(u, u)}{u},$$

where $\bar{C}$ is the survival copula defined by $\bar{C}(1-u, 1-v) = 1-u-v+C(u, v)$.

5.8 Summary

For discrete state spaces the Markov property is written as

$$P\left[X_t = a \middle| X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P[X_t = a | X_s = x],$$

for all $s_1 < s_2 < \cdots < s_n < s < t$ and all states $a, x_1, x_2, \ldots, x_n, x$ in $\mathcal{S}$.

Any process with independent increments has the Markov property.

Markov chains are discrete-time and discrete-state-space stochastic processes satisfying the Markov property. You should be familiar with the simple NCD, modified NCD, unrestricted random walk and restricted random walk processes.

In general, the n-step transition probabilities $p_{ij}(m, m+n)$ denote the probability that a process in state i at time m will be in state j at time $m+n$.

The transition probabilities of a Markov process satisfy the Chapman–Kolmogorov equations:

$$p_{ij}(m, n) = \sum_{k \in \mathcal{S}} p_{ik}(m, l) p_{kj}(l, n),$$

for all states $i, j \in \mathcal{S}$ and all integer times $m < l < n$. This can be expressed in terms of n-step stochastic matrices as

$$\mathbf{P}(m, n) = \mathbf{P}(m, l)\mathbf{P}(l, n).$$

An irreducible time-homogeneous Markov chain with a finite state space has a unique stationary probability distribution, π, such that

$$\pi = \pi \mathbf{P}^{(n)}.$$

Aperiodic processes will converge to the stationary distribution as $n \to \infty$.

6.6 Summary

Markov jump processes are continuous-time and discrete-state-space stochastic processes satisfying the Markov property. You should be familiar with the Poisson, survival, sickness-death and marriage models.

The Poisson process is a simple Markov jump process. It is time-homogeneous with stationary increments that are Poisson distributed with mean $\lambda > 0$. Waiting times between jumps are exponentially distributed with mean $1/\lambda$.

As with Markov chains, transition probabilities exist for a general Markov jump process

$$p_{ij}(s,t) = P\left[X_t = j | X_s = i\right], \text{ where } p_{ij}(s,t) \geq 0 \text{ and } s < t,$$

which must also satisfy the Chapman-Kolmogorov equations.

The quantities $q_{jj}(t), q_{kj}(t)$ are the transition rates, such that

$$\begin{aligned} \lim_{h \to 0} \frac{p_{jj}(t,t+h) - 1}{h} &:= q_{jj}(t), \\ \lim_{h \to 0} \frac{p_{kj}(t,t+h)}{h} &:= q_{kj}(t), \quad \text{for} \quad k \neq j. \end{aligned}$$

Kolmogorov's forward and backwards equations are respectively

$$\frac{\partial p_{ij}(s,t)}{\partial t} = \sum_{k \in \mathcal{S}} p_{ik}(s,t) q_{kj}(t) \quad \text{and} \quad \frac{\partial p_{ij}(s,t)}{\partial s} = -\sum_{k} q_{ik}(s) p_{kj}(s,t).$$

These can be written in matrix form as

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}(t) \quad \text{and} \quad \frac{\partial \mathbf{P}(s,t)}{\partial t} = -\mathbf{Q}(s)\mathbf{P}(s,t),$$

where $\mathbf{Q}(t)$ is the generator matrix with entries $q_{ij}(t)$.

In time-homogeneous models the time-dependence of the transition probabilities and transition rates (therefore generator matrices) is removed.

The residual holding time R_s is the random amount of time between time s and the next jump:

$$\{R_s > w, X_s = i\} = \{X_u = i, s \leq u \leq s + w\}.$$

It can be proved that

$$P(R_s > w |\, X_s = i) = e^{\int_s^{s+w} q_{ii}(t) dt},$$

.

7.6 Summary

Machine learning can be defined as the study of systems and algorithms that improve their performance with experience.

Machine learning typically used to solve classification problems, clustering, regression problems, analysis of association rules, discovering hidden or latent variables, etc.

One of the main problems in machine learning is overfitting, when the system works perfectly on the data it was trained in but performs badly on any other data. To test that it did not occur, one may divide our data into two groups, and use one group for training, and another one for testing. We can then exchange the role of training and testing data, this procedure is called cross validation. In fact, data for training can be further divided into two groups: a training data set and validation data set, the first one is to find model parameters, and the second one is to find hyper-parameters. The training-validation-test data proportion may be, for example, $60\% - 20\% - 20\%$.

In the yes-no classification problem, the correct outcomes are true positive (TP), and true negative (TN), while the incorrect ones are False Positive (FP) and False negative (FN). This can be used to calculate various measures of performance, such as Precision, Recall, F1 score, and False Positive Rate.

You should be able to define and understand the following terms and methods:

- Supervised learning
- Unsupervised learning
- Linear classifier
- Support vector machines
- Nearest-neighbour classifier
- K-means algorithm
- Maximal a posteriori (MAP) decision rule
- Decision tree
- Likelihood function
- Naive Bayes classifier
- Reinforcement learning

Machine Learning tasks can often be broken down into a series of steps:

- Collecting data
- Exploring and preparing the data
- Feature scaling
- Splitting the data into the training, validation and testing data sets
- Training a model on the data
- Validation and testing
- Improving model performance
- The reproducibility of research