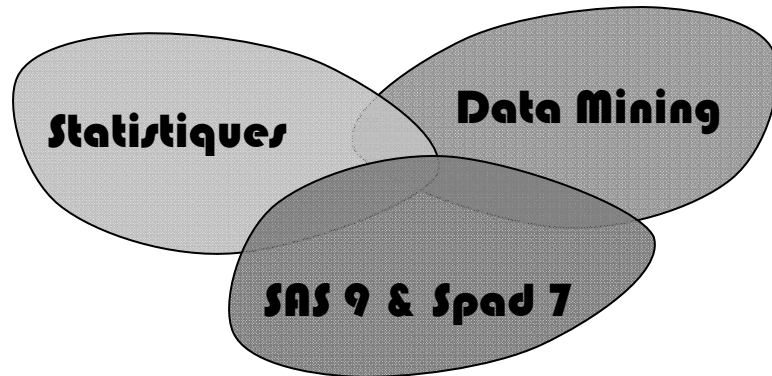


Olivier Decourt SARL



Contenu des cours 2012

Nous contacter par email à :
formation@od-datamining.com

Visitez notre site : <http://www.od-datamining.com/>

Sommaire

| | |
|---|----|
| Tarifs et conditions | 3 |
| Formations statistiques | 4 |
| [AASAT] Statistique descriptive avec SAS | 5 |
| [ANADON] Analyse des données sous SAS | 6 |
| [ANOVA] Analyse de la variance | 7 |
| [BOOT] Le bootstrap, théorie et pratique | 8 |
| [GENMOD] Modèle linéaire généralisé | 9 |
| [IMPUTER] Traitement de la non-réponse, calage & imputation | 10 |
| [MIXED] Analyse de la variance et modèles mixtes | 11 |
| [NEWSTAT9] Nouveautés de SAS/STAT en version 9 | 12 |
| [ODSGRAPH9] ODS GRAPHICS | 13 |
| [PLS] La régression PLS | 14 |
| [REG] Techniques de régression | 15 |
| [REGQUALI] Régression sur variables qualitatives | 16 |
| [REGQUANTI] Régression sur variables quantitatives | 17 |
| [STATD] De la description à la décision : initiation à SAS/STAT | 18 |
| [STAT_SEG] Statistiques avec SAS Enterprise Guide | 19 |
| [SURVIE] Analyse de survie et économétrie des durées | 20 |
| Formations au DataMining | 21 |
| [DM] Qu'est-ce que le Data Mining ? | 22 |
| [SCORING] Panorama et comparaison des méthodes de scoring | 23 |
| [RN] Les réseaux de neurones | 24 |
| [ARBRES] Les arbres de décision | 25 |
| [SEM4] SAS Enterprise Miner version 4 | 26 |
| [SEM5] SAS Enterprise Miner version 5 et 6 | 27 |

Tarifs et conditions

Groupes de 1 à 8 personnes

Nous proposons un tarif unique en intra-entreprise, quel que soit le niveau du cours : 1 200 € HT par jour de formation. Ce tarif est applicable à tout groupe de huit personnes maximum. Il inclut la réalisation et la distribution de supports de cours pour chaque participant.

Des formations sur mesure

Les contenus de cours qui sont détaillés dans les pages suivantes ne sont pas limitatifs. Il ne s'agit que des contenus standard. Ceux-ci peuvent être adaptés à vos besoins, qu'il s'agisse de la modification d'un contenu existant ou de la création d'un nouveau cours. Ces modifications n'entraînent aucune modification des tarifs ci-dessus.

Cours en province

Les frais de déplacement et d'hébergement sur place du formateur seront facturés en sus sur présentation de justificatifs.

Formations statistiques

[AASSTAT] Statistique descriptive avec SAS

Ce stage est destiné aux personnes désireuses de (re)découvrir les principes de la statistique exploratoire. La mise en œuvre de ces techniques se fait autour des procédures de SAS/GRAPH et SAS/STAT.

Ce cours est une bonne préparation aux formations de modélisation (REGQUANTI, REGQUALI, GENMOD) ainsi qu'à l'analyse de la variance (ANOVA).

Durée : 2 jours

• DECRIRE LES DONNEES PAR DES GRAPHIQUES

- Graphiques univariés (bâtons et diagrammes circulaires) avec la procédure GCHART
- Graphiques bivariés (nuages de points et boîtes à moustaches) avec les procédures GPLOT et BOXPLOT
- Graphiques de répartition d'une variable avec les procédures UNIVARIATE et KDE

• INDICATEURS STATISTIQUES USUELS

- Rappels sur les définitions des moyennes, médianes, variances, etc.
- Calcul de statistiques descriptives avec la procédure MEANS
- Calcul de statistiques plus poussées avec la procédure UNIVARIATE

• RECHERCHE DE LIAISONS ENTRE VARIABLES

- Tableaux de fréquences, chi-2 avec la procédure FREQ
- Corrélations linéaires avec la procédure CORR
- Non-linéarité des liaisons et transformation quanti/quali : l'utilité d'un format

• TESTS STATISTIQUES

- Principe d'un test statistique
- Test d'adéquation à une loi avec la procédure UNIVARIATE
- Test du chi-2 et quantités dérivées avec la procédure FREQ
- Test de comparaison de moyennes et de proportions avec la procédure TTEST

• INTRODUCTION A LA MODELISATION

- Principe d'un modèle statistique
- Régression linéaire avec la procédure GLM : premiers résultats
- Vérification des hypothèses du modèle : de l'importance d'une bonne analyse exploratoire
- Tableau synthétique des modèles disponibles dans le module SAS/STAT

[ANADON] Analyse des données sous SAS

Ce stage est destiné aux chargés d'études qui désirent voir ou revoir les principes de l'analyse de données à la française (ACM, AFC, ACP) et surtout leur utilisation à travers SAS. On y aborde également la classification, avec SAS et SAS Enterprise Miner.

Durée : 2 jours

- **L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)**
 - Syntaxe de la proc PRINCOMP
 - Choix du nombre d'axes factoriels
 - Nuages des individus et des variables
 - Cercle des corrélations

- **L'ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)**
 - Création d'un tableau disjonctif complet
 - Syntaxe de la proc CORRESP
 - Choix du nombre d'axes factoriels
 - Nuages des individus et des variables
 - Individus et variables supplémentaires

- **TYPLOGIES**
 - Classification ascendante hiérarchique avec la proc CLUSTER
 - Nuées dynamiques avec la proc FASTCLUS
 - Méthode mixte de Wong
 - Description des classes
 - Modéliser l'appartenance aux classes pour réaffecter

[ANOVA] Analyse de la variance

Avec des bases statistiques (description univariée : moyenne, variance, quantiles), ce stage propose de découvrir des outils puissants : l'analyse de la variance et les tests statistiques d'égalité de moyennes.

Durée : 2 jours

- **L'ANALYSE DE LA VARIANCE ET LE MODELE LINEAIRE GENERAL**
 - Hypothèses du modèle linéaire général
 - Validation des hypothèses
 - Décomposition de la variance et tests afférents
 - L' « analyse de la déviance » et les MLG

- **LES TESTS DE COMPARAISONS DE MOYENNES**
 - Tests simultanés
 - Test à un groupe de référence : test de Dunnett
 - Contrastes
 - Moyennes ajustées ou non

- **L'ETUDE DES DONNEES REPETEES ET CORRELEES**
 - Cas de mise en œuvre
 - Dans le cadre du modèle linéaire général
 - Dans le cadre du modèle linéaire généralisé

[BOOT] Le bootstrap, théorie et pratique

Ce cours propose de comprendre les concepts du bootstrap et du jackknife, et de permettre leur mise en oeuvre dans les processus d'étude et de modélisation. La connaissance de la programmation SAS et du macro-langage est requise.

Durée : 1 jour

- **QUELQUES DEFINITIONS**

- Méthode de Monte-Carlo
- Echantillonnage et rééchantillonnage
- Plug-in

- **LE BOOTSTRAP POUR LA CONSTRUCTION D'INTERVALLES DE CONFIANCE**

- Bootstrap T
- Bootstrap percentile
- Bootstrap BCa

- **LE BOOTSTRAP POUR LES PROBLEMES DE CLASSIFICATION**

- Bootstrap .632
- Bagging
- Boosting

- **MISE EN ŒUVRE SOUS SAS**

- Bootstrap
- Bagging
- Boosting

[GENMOD] Modèle linéaire généralisé

Les modèles présentés ici font de la régression linéaire et de la régression logistique des cas particuliers. Les Modèles Linéaires Généralisés (MLG) se proposent d'étudier les variables dont la normalité est prise en défaut (coûts, fréquences d'évènements, ...) et proposent des outils puissants.

Durée : 2 jours

- **PRINCIPES DE LA REGRESSION**

- Vocabulaire et concepts
- La régression linéaire
- La régression logistique
- Leurs points communs

- **MODELE LINEAIRE GENERALISE**

- Loi de Y
- Fonction de lien
- Qualité du modèle
- Analyse de la déviance
- Analyse des résidus et autres vérifications
- Syntaxe de la procédure GENMOD de SAS

- **EXEMPLES DE MODELES LINEAIRES GENERALISES**

- Régression de Poisson
- Régression binomiale négative
- Régression Gamma
- Données répétées ou corrélées : les GEE

[IMPUTER] Traitement de la non-réponse, calage & imputation

Les résultats d'une enquête sont souvent entachés de non-réponse : des individus n'ont, partiellement ou totalement, pas répondu aux questions qui leur étaient posées.

Cette formation orientée vers la pratique sous SAS des techniques de redressement usuelle permet, en une journée, de balayer les différents types de non réponse et les solutions que l'on peut y apporter.

Durée : 1 jour

- **NON REPONSE TOTALE : CALAGE**

- Le problème initial
- Pourquoi modifier les poids
- Quelles variables de calage ?
- Le calage sur marges : théorie
- Le calage sur marges : pratique avec la macro %CALMAR

- **NON REPONSE PARTIELLE : IMPUTATION**

- Imputation par la moyenne
- Imputation par régression
- Imputation par hotdeck
- Imputation par résidus simulés
- Imputation par plus proches voisins

[MIXED] Analyse de la variance et modèles mixtes

L'étude des données avec une analyse de la variance se conduit d'ordinaire sur des facteurs considérés comme fixes : c'est à dire qu'on se limite dans l'analyse et l'inférence aux valeurs qui ont été collectées au cours de la constitution des données. Des facteurs aléatoires et un modèle mixte étendent de manière très importante la puissance des modèles d'analyse de variance, et facilitent également le traitement des données à mesures répétées

Durée : 2 jours

- **ANALYSE DE VARIANCE, EFFETS FIXES ET ALEATOIRES**
 - Buts et hypothèses de l'analyse de variance
 - Effet fixe et effet aléatoire
 - Théorie et notations
 - Panorama de l'offre SAS pour l'analyse de variance

- **ANALYSE DE LA VARIANCE A EFFETS ALEATOIRES**
 - Syntaxe de la procédure MIXED
 - Détection graphique d'effets
 - Quantification d'un effet aléatoire, calcul de moyennes ajustées
 - Comparaison de groupes, ajustements pour les comparaisons multiples
 - Intégration de variables fixes quantitatives
 - Interactions

- **MODELES MIXTES GENERALISES**
 - Principe et théorie des modèles linéaires généralisés
 - Syntaxe de la procédure GLIMMIX
 - Régression logistique à effets aléatoires
 - Régression de Poisson à effets aléatoires
 - Régression Gamma à effets aléatoires

- **ANALYSE DE VARIANCE SUR DONNEES REPETEES**
 - Variabilité inter-sujets et intra-sujets
 - Les principales structures de covariance
 - Comparaison et choix de la structure la plus adaptée aux données

[NEWSTAT9] Nouveautés de SAS/STAT en version 9

En une journée, découvrez et pratiquez les principales nouveautés du module SAS/STAT en version 9 de SAS.

Durée : 1 jour

- **L'ODS ET LES STATISTIQUES**

- Fonctionnement général de l'ODS
- L'ODS et les tables SAS
- ODS GRAPHICS
- Le GTL (Graph Template Language)

- **NOUVEAUTES GRAPHIQUES**

- Proc BOXPLOT
- Proc UNIVARIATE

- **NOUVEAUTES POUR LA PREPARATION DES DONNEES**

- Proc KDE
- Procs MI et MIANALYZE
- Proc STDIZE
- Proc SURVEYSELECT

- **NOUVEAUTES POUR LA MODELISATION**

- Proc GENMOD
- Procs GLM et REG
- Proc LOGISTIC
- Proc PLS
- Proc ROBUSTREG
- Proc GLIMMIX
- Proc GLMSELECT

- **NOUVEAUTES DIVERSES**

- Proc FASTCLUS
- Procs SURVEY...
- Proc FREQ
- Proc TTEST

[ODSGRAPH9] ODS GRAPHICS

Cette formation concerne les utilisateurs de SAS voulant comprendre et tirer avantage d'ODS GRAPHICS, le système d'édition automatique de graphiques statistiques de SAS version 9. Cette formation propose également une initiation aux procédures SGPLOT et SGPANEL, et à la production de graphiques sur mesure avec le langage GTL au cœur d'ODS Graphics.

Durée : 2 jours

- **PRINCIPE DE L'ODS - RAPPELS**

- Objets
- Style global
- Style tabulaire
- Rôle de la PROC TEMPLATE

- **UTILISATION D'ODS GRAPHICS**

- Destinations ODS concernées
- Procédures supportant ODS GRAPHICS
- Exemples de syntaxe
- Où trouver la documentation sur ODS GRAPHICS ?

- **UTILISATION AVANCEE D'ODS GRAPHICS**

- Graphiques « actifs » et mapping
- Sauvegarde d'un graphique en fichier séparé
- Dégroupage de graphiques
- Utilisation de la PROC TEMPLATE

- **PROCEDURES SGxxx**

- Procédure SGPLOT : graphiques sur variables qualitatives
- Procédure SGPLOT : graphiques sur variables quantitatives
- Procédure SGPANEL : une image, plusieurs graphiques

- **GRAPH TEMPLATE LANGUAGE ou GTL**

- Fonctionnement des templates GTL
- Utilisation d'un template GTL (via une étape Data, une procédure statistique, la procédure SGRENDER)
- Syntaxe du GTL, construction de graphiques à façon

[PLS] La régression PLS

Une méthode « moderne » qui acquiert lentement mais sûrement une certaine notoriété. C'est un outil nécessaire dans beaucoup de domaines où la redondance de l'information est un problème épineux : chimie, physique, sociologie, économie, ...

Durée : 1 jour

- **LE PROBLEME INITIAL**

- Régression sur données corrélées
- Les symptômes de la multicolinéarité
- Les remèdes usuels et leurs limites

- **RESUMER L'INFORMATION**

- L'analyse en composantes principales
- L'analyse des correspondances multiples
- L'analyse canonique des corrélations
- Quelle solution ?

- **LA REGRESSION « PLS1 »**

- Construction des composantes
- Choix du nombre de composantes
- Résultats usuels de la régression PLS1

- **MISE EN ŒUVRE SOUS SAS**

- La procédure PLS
- Comparaison avec une régression sur axes factoriels
- Quelques programmes pour compléter les sorties de la PROC PLS

[REG] Techniques de régression

Une formation complète qui propose aux chargés d'études un maximum de solutions de modélisation linéaire pour données de tous types : continues, catégorielles, binaires.

Durée : 3 jours

- **QU'EST-CE QU'UN MODELE LINEAIRE ?**

- Les régressions usuelles
- Les hypothèses du modèle linéaire
- Principe des tests statistiques

- **LE MODELE LINEAIRE CLASSIQUE**

- Hypothèses et validation des hypothèses
- Qu'est-ce qu'un modèle réussi ?
- Les coefficients et leurs p-values
- Les résidus
- Exemple « criminalité »

- **ANALYSE DE LA VARIANCE, MODELE LINEAIRE GENERAL**

- Hypothèses et validation des hypothèses
- Introduction de facteurs qualitatifs
- Analyse de variance : lien avec le modèle linéaire général
- Comparaison de moyennes
- Moyennes ajustées (LSMEANS)
- Exemple « éducation »

- **REGRESSION LOGISTIQUE**

- Critères de qualité du modèle (Akaike, Schwarz)
- Les coefficients et les odds-ratios
- Notion de score, aide à la décision (seuil optimal)
- Courbe ROC
- Exemple du Titanic
- Exemple sur l'assurance automobile

- **MODELE LINEAIRE GENERALISE**

- Lois autorisées dans un tel modèle
- Fonction de lien
- Loi des résidus
- Qualité du modèle
- Analyse de la déviance
- Régression de Poisson
- Régression Gamma

[REGQUALI] Régression sur variables qualitatives

Destiné aux chargés d'étude s'intéressant à la modélisation d'une variable discrète (deux modalités ou davantage), ce stage permet de construire efficacement des modèles explicatifs et prédictifs (construction de scores).

Durée : 2 jours

- **PRINCIPE DE LA REGRESSION LOGISTIQUE**

- Quelle est la forme des données à utiliser ?
- Lien avec la régression linéaire
- Les différentes fonctions de lien
- Mesurer la qualité d'un modèle logistique
- Syntaxe de base de la procédure Logistic de SAS
- Exemple des maladies coronariennes

- **LA REGRESSION LOGISTIQUE A BUT DESCRIPTIF**

- L'analyse de la déviance, étude de l'impact d'une covariable
- Stratégies de construction de modèles cohérents
- Les coefficients
- Les odds-ratios
- La multicolinéarité
- Exemple du Titanic

- **LA REGRESSION LOGISTIQUE A BUT PREDICTIF**

- Qu'est-ce qu'un score ?
- La courbe ROC et le seuil optimal
- La courbe de lift
- Qualité d'ajustement
- Syntaxe de la procédure Logistic pour la prédiction
- Exemple sur l'assurance automobile

- **ETUDE D'UNE VARIABLE A PLUSIEURS MODALITES**

- Régression sur une variable ordonnée
- Régression sur une variable non ordonnée ou logit généralisé
- Application à la description d'une typologie

- **MODELISATIONS ALTERNATIVES D'UNE VARIABLE QUALITATIVE**

- Analyse discriminante
- Réseaux de neurones
- Arbres de décision

[REGQUANTI] Régression sur variables quantitatives

Ce cours permet d'appréhender les principes de la régression, et sa mise en oeuvre sous SAS (procédures REG et GLM). On y apprend le formalisme statistique associé, mais surtout la lecture des résultats, la détection d'erreurs et leur correction.

Durée : 2 jours

- **DECOUVERTE DES DONNEES**
 - Distribution et normalité des variables
 - Relations entre variables quantitatives
 - Relations entre variables qualitatives

- **REGRESSION LINEAIRE SIMPLE**
 - Le modèle simple
 - Sorties chiffrées
 - Sorties graphiques

- **SELECTION D'UN MODELE OPTIMAL**
 - Méthodes pas à pas
 - Sélection sur un critère

- **COMBATTRE LA MULTICOLINEARITE**
 - Détecter la multicolinéarité
 - Régression sur composantes factorielles
 - Régression PLS

- **GESTION DES COVARIABLES QUALITATIVES**
 - Le modèle linéaire général
 - Choix de la modalité de référence
 - Lecture des sorties de la procédure GLM

[STATD] De la description à la décision : initiation à SAS/STAT

Ce cours se propose de faire découvrir les principales fonctionnalités offertes par SAS dans le domaine du décisionnel (processus d'étude statistique ou DataMining). Il s'adresse aux chargés d'études statistiques, ainsi qu'aux personnes ayant à mener des projets de DataMining sans progiciel spécifique.

Durée : 3 jours

- **LA DESCRIPTION DES DONNEES**

- La proc MEANS
- La proc BOXPLOT
- La proc UNIVARIATE

- **LES INTERACTIONS**

- La proc FREQ
- La proc CORR
- La proc TTEST

- **LES METHODES FACTORIELLES**

- La proc PRINCOMP
- La proc CORRESP

- **LA CLASSIFICATION**

- La proc CLUSTER
- La proc FASTCLUS
- La méthode mixte

- **MODELISER**

- La proc REG
- La proc GLM
- La proc LOGISTIC
- La proc DISCRIM

[STAT_SEG] Statistiques avec SAS Enterprise Guide

Ce cours se propose de faire découvrir les principales fonctionnalités de statistique descriptives via les écrans de dialogue presse-bouton de SAS Enterprise Guide. Il s'adresse aux chargés d'études statistiques ne souhaitant pas coder de programmes SAS.

Durée : 2 jours

- **ANALYSE DESCRIPTIVE**

- Calcul de statistiques de base : moyenne, médiane, écart-type
- Tableaux de fréquences
- Adéquation à une loi statistique
- Rappels sur les principaux tests
- Croisement de variables qualitatives et tests du χ^2
- Comparaison de moyennes à 2 groupes (test de Student)
- Comparaison de moyennes à k groupes (analyse de la variance)

- **ANALYSE FACTORIELLE, SEGMENTATION**

- ACP pour les variables quanti
- Segmentation / classification par méthode ascendante hiérarchique
- Segmentation par méthode des K-moyennes

- **MODELISATION**

- Régression linéaire simple et multiple
- Analyse de variance
- Régression logistique

[SURVIE] Analyse de survie et économétrie des durées

Ce stage est destiné aux personnes ayant à étudier la durée écoulée avant la survenance d'un évènement. Il s'agit par exemple d'un contexte médical (durée de rémission dans une maladie chronique) ou économique (durée de recherche d'emploi)... Cette formation propose à la fois une présentation théorique (avec un formalisme mathématique aussi léger que possible) et appliquée sous SAS (procédures LIFETEST, PHREG et LIFEREG).

Durée : 1 jour

- **DONNEES CENSUREES, COMMENT LES DECRIRE**

- Notion de censure
- Modèles pour données censurées
- Fonction de survie
- Quartiles
- Intervalles de confiance
- Méthode de Nelson-Aalen
- Comparaison de groupes
- Autres représentations graphiques des données

- **MODELE DE COX**

- Hasard instantané, hasard cumulé
- Ratio de hasards
- Modèle semi-paramétrique
- Hypothèses à vérifier – et comment on les vérifie
- Le hazard-ratio, une notion-clé
- Sélection de variables
- Prédications

- **MODELE PARAMETRIQUE « ACCELERE »**

- Quelle loi choisir dans un modèle paramétrique ?
- Interprétation des sorties de la procédure LIFEREG pour les modèles de durée/survie
- Comparaison des sorties entre la proc LIFEREG et la proc PHREG

Formations au DataMining

[DM] Qu'est-ce que le Data Mining ?

Une formation destinée aux chargés de projets et aux décideurs qui veulent savoir ce que recouvre exactement le mot de Data Mining. Quels sont les concepts, les démarches, les outils méthodologiques, les logiciels du marché avec leurs forces et leurs faiblesses ?

Durée : 1 jour

- **DEFINITION DU DATA MINING**

- Un peu d'histoire
- Les domaines "historiques" d'application
- De nouveaux domaines d'expression

- **LES TECHNIQUES DU DATA MINING**

- La méthodologie
- Les arbres de décision
- Les réseaux de neurones
- Les raisonnements à base de cas (MBR)
- Les machines à vecteurs-supports (SVM)
- Qu'est-ce qu'un score ?

- **L'OFFRE LOGICIELLE**

- Les prérequis
- Les critères importants
- Quelques outils comparés

[SCORING] Panorama et comparaison des méthodes de scoring

Cette formation s'adresse aux chargés d'étude désirant avoir, en quelques jours, un aperçu technique et pratique (avec des logiciels comme SAS/SEM, SPAD ou SPSS/Clementine/Answer Tree) des techniques usuelles de scoring. La formation s'achève avec une comparaison des forces et des faiblesses des différentes méthodes.

Durée : 3 jours

- **Scoring avec les arbres de décision**
 - Principe général d'un arbre de décision
 - Croissance et élagage
 - Les principaux algorithmes : CHAID, CART, C4.5
 - Arbres, bagging et boosting : comment rendre un arbre robuste
 - Avantages et inconvénients

- **Scoring avec la régression logistique**
 - Modèle linéaire et modèle logistique
 - Choix des variables, automatisation
 - Coefficients et odds-ratios
 - Courbe ROC, discrimination
 - Avantages et inconvénients

- **Scoring avec l'analyse discriminante**
 - Approche géométrique
 - Fonction linéaire discriminante
 - Méthode DISQUAL : l'analyse discriminante sur données qualitatives
 - Avantages et inconvénients

- **Scoring avec les réseaux de neurones**
 - Le neurone artificiel
 - Apprentissage supervisé
 - Lecture et interprétation des résultats
 - Avantages et inconvénients

- **Autres méthodes de scoring**
 - Raisonnement basé sur la mémoire
 - Machines à vecteurs supports (méthode Vapnik)
 - Bagging et boosting

- **Comparaison générale des méthodes de scoring**

[RN] Les réseaux de neurones

Cette formation propose de découvrir les principes et les applications des réseaux de neurones, comment les mettre au point, comment interpréter leurs résultats et comment faire le choix du meilleur réseau.

Durée : 1 jour

- **PRINCIPES DES RESEAUX DE NEURONES**

- Le modèle humain et le neurone artificiel
- Les fonctions de transfert
- Les couches et les liaisons synaptiques
- Intérêt statistique de la démarche : les réseaux de neurones comme modèles
- L'apprentissage et le sur-apprentissage

- **LES PERCEPTRONS**

- Principe
- Avantages et inconvénients

- **LES RESEAUX RBF**

- Principe
- Avantages et inconvénients

- **LES CARTES AUTO-ORGANISEES**

- Les cartes de Kohonen
- Intérêt statistique : une autre lecture des classifications SOM

[ARBRES] Les arbres de décision

Présentation des grands principes et mise en œuvre des arbres de décision avec différents logiciels : SAS Enterprise Miner, Answer Tree, Spad. Les arbres de décision sont à la fois un outil décisionnel, avec une optique de modélisation, et un outil exploratoire, pour la préparation et la découverte des données.

Durée : 1 jour

- **GENERALITES**

- Vocabulaire
- Méthodologie de construction
- Offre logicielle
- Utilisations d'un arbre

- **DEVELOPPEMENT D'UN ARBRE**

- Choix de la coupure
- Impact du critère d'évaluation de la coupure
- Critères d'arrêt

- **ELAGAGE ET VALIDATION**

- Cas de l'algorithme CHAID
- Elagage
- Evaluation des sous-arbres

- **AMELIORATION D'UN ARBRE DE DECISION**

- Evolutions méthodologiques
- Bagging
- Boosting

- **FORCES ET FAIBLESSES**

[SEM4] SAS Enterprise Miner version 4

Prise en main de la solution DataMining de SAS : méthodologie, principales fonctionnalités. Ce cours s'appuie sur des exemples concrets de création de typologies et de mise au point d'un moteur de score.

Durée : 2 jours

- **LE DATAMINING, QUI, COMMENT, POURQUOI ?**
 - Qui ?
 - Comment ?
 - Pourquoi ?

- **PREMIER CONTACT AVEC SAS ENTERPRISE MINER**
 - Démarrage de SEM
 - Notions de projet et de diagramme
 - Les projets

- **APPREHENSION ET MISE EN CONFORMITE DES DONNEES**
 - Insérer des données dans le diagramme
 - Echantillonnages
 - Galerie de graphiques
 - Graphiques à façon
 - Choix des variables retenues pour l'étude
 - Gestion des valeurs manquantes
 - Filtres sur les individus

- **MODELISATION ET SCORE**
 - Les modèles
 - La régression
 - Les arbres de décision
 - Les réseaux de neurones : le perceptron
 - Mise en concurrence et comparaison de modèles
 - Le nœud SCORE : produire une étape Data
 - Après le nœud SCORE, plus rien à faire ?
 - Le nœud REPORTING et les rapports en HTML

- **SEGMENTATION**
 - Un « nouveau » diagramme
 - Les K-moyennes
 - Interprétation des classes
 - Conclusion sur la construction d'une typologie

[SEM5] SAS Enterprise Miner version 5 et 6

Les versions 5 et 6, conçues comme des clients légers Java, de SAS EM sont assez différentes des précédentes. Selon qu'il s'agit ou non de votre première expérience avec Enterprise Miner, la formation dure une ou deux journées.

Durée : 1 jour (montée de version) à 2 jours (initiation)

- **PREMIER CONTACT AVEC SAS ENTERPRISE MINER**

- Notions de projet, de diagramme et de noeud
- Démarrage de SEM
- Ecran d'accueil
- Les sources de données
- Créer un nouveau diagramme
- Explorer les données – édition liminaire de graphiques

- **PHASE EXPLORATOIRE**

- Insérer des données dans le diagramme
- Echantillonnages
- Exploration numérique des données
- Exploration graphique des données
- Filtres sur les individus
- Gestion des valeurs manquantes

- **MODELISATION ET SCORE**

- Généralités sur la modélisation
- Evaluation d'un modèle de score
- Les arbres de décision
- La régression
- Les réseaux de neurones : le perceptron
- Mise en concurrence et comparaison de modèles
- Le noeud SCORING : produire un moteur de score
- La nécessaire intégration des scores au Data Warehouse

- **SEGMENTATION**

- Un « nouveau » diagramme
- Principes de la classification mixte
- Construction de classes
- Les sorties de la troisième segmentation
- Conclusion sur la construction d'une typologie