Forecasting stock market movement direction with support vector machine

John Sibony

15 Mars 2018

Ce rapport détaille seulement une partie des analyses et méthodes explicitées dans le Notebook Jupiter. Ce dernier est donc à consulter de préférence antérieurement.

Introduction

La prédiction des marchés financiers est une tâche très complexe, car les séries financières sont intrinsèquement bruyantes, non stationnaires et chaotiques. La caractérisation de bruit se réfère à l'indisponibilité d'informations complètes sur le comportement passé des marchés financiers pour saisir pleinement la dépendance entre les prix futurs et passés. Les informations manquantes non incluses dans notre modèle sont considérées comme du bruit. Par exemple à l'inverse de la reconnaissance d'objets où les caractéristiques de détection sont explicitement connus, les séries financières évoluent selon un panel de variables inconnus. La caractéristique non stationnaire implique que la distribution des séries évolue avec le temps : la dynamique des prix dépend de la variable temporelle. Par chaotique, on entend que les séries chronologiques financières sont aléatoires à court terme mais déterministes à long terme. De nombreux facteurs et événements inattendus tels que la situation économique et politique ou les attentes des opérateurs peuvent entraîner un changement d'une série temporelle financière. Dans le même temps, la relation de toute série financière avec les autres séries de données connexes peut également changer avec le temps. Par conséquent, prédire les mouvements du marché financier est assez difficile et instable.

La négociation des indices boursiers représente une part importantes des volumes trader sur les principaux marchés. La prévision précise des ces indices est importante pour de nombreuses raisons. Tout d'abord figure la nécessité pour les investisseurs de se prémunir contre les risques de marché potentiels, mais également les opportunités pour les spéculateurs et les arbitragistes de réaliser des profits en négociant des indices boursiers. De toute évidence, être capable de prévoir avec l'indice boursier a de profondes implications et une signification pour

les chercheurs et praticiens.

Il existe de beaucoup de littératures qui se concentrent sur la prévisibilité des valeurs des indices boursiers. Dans presque tous les cas, les mesures de performance et l'acceptabilité des modèles proposés sont mesurées par les écarts de la valeur de prévision par rapport aux valeurs réelles. Différents investisseurs adoptent des stratégies de négociation différentes, de sorte que les modèles de prévision qui se classent d'abord en termes de minimisation de l'erreur de prévision peuvent ne pas convenir à ces acteurs. Cependant, certaines études ont suggéré que les stratégies de négociation guidées par des prévisions sur la direction de la variation des prix pourraient être plus efficaces et générer des profits plus élevés (Leung et al., 2000). De plus, prédire la direction est une question pratique qui affecte généralement la décision d'un opérateur financier d'acheter ou de vendre un instrument.

Par conséquent notre étude s'est porté sur la prédiction de la direction du Nikkei 225, principal indice boursier de la bourse de Tokyo, en se basant sur l'article de recherche "Forecasting stock market movement direction with support vector machine" (Wei Huang et al, 2004). Différents algorithmes tels que le SVM, la LDA et la QDA ont été utilisés et comparés à travers la principale mesure de performance en classification binaire à savoir le hit ratio (proportion de classification correcte). Nous proposons tout d'abord dans ce rapport une revue de la littérature des diverses techniques de classification liées à la prévision des séries chronologiques. Puis nous détaillerons les algorithmes utilisés ainsi que les différentes techniques apportées pour améliorer la robustesse des classifieurs. Enfin, nous analyserons et détaillerons les résultats obtenus par notre propre modèle en adoptant une comparaison critique avec l'article.

I - État de l'art des méthodes de prédiction

1 - Efficience des marchés

Avant de poursuivre notre étude, il serait intéressant de se demander si notre projet de prédiction du marché est viable. L'hypothèse d'efficience du marché induit que aucun investisseur ne peut réussir à obtenir un profit anormal sur le marché pour un certain niveau de risque donné. Sur le long terme, battre le marché est donc impossible. Le prix d'un actif est donc égal à sa valeur théorique et la surévaluation ou sous-évaluation d'actif n'a pas de sens dans un marché efficient. Il existe 3 degrés plus ou moins important d'efficience des marchés : la forme faible, semi-forte et forte. La première consiste à supposer que les données passés de l'indice sont déjà incorporées par le prix actuel du cours. L'analyse du cours historique est alors inutile. La seconde forme implique qu'en plus des historiques du cours, toute l'information publique et est incorporée dans les prix de l'indice. L'analyse d'indicateur économique est alors rendu obsolète. Enfin la dernière forme suppose en plus que l'information relevant du domaine privé est également reflétée dans les prix actuels. Ainsi, l'efficience des marché rend compte du degré d'omniscience des prix.

La théorie d'efficience des marchés divise les chercheurs. Cependant, le nombre considérable de littérature sur la prédiction des cours financiers démontre que les rendements boursiers sont dans une certaine mesure prévisibles. Pour les États-Unis, Fama et French ont étudiés l'impact de certaines variables financières comme puissance de prédiction des rendements boursiers. Des études basées sur les marchés

européens rapportent des résultats similaires. Les résultats de Ferson et Harvey indiquent que les rendements sont, dans une certaine mesure, prévisibles sur un certain nombre de marchés européens (Royaume-Uni, France, Allemagne). Dans leur étude visant à prévoir les cours des actions au Royaume-Uni, Jung et Boyd font état d'une performance raisonnablement bonne de leurs prévisions, suggérant que la force prédictive de leurs modèles de rendement boursier n'est pas négligeable. Enfin, pour le marché boursier japonais, les recherches empiriques de Jaffe et Westerfield (1985) et Kato et al. (1990) ont également prouvés a prévisibilité du comportement des indices.

2 - Approche technique et fondamentale

Les caractéristiques (input) utilisées par les classifieurs permettant la prédiction des indices (output) peuvent être classifiées en une approche technique ou fondamentale. L'analyse technique statistique repose sur la configuration même du cours de l'indice en question. L'historique des prix est analysé à travers des capteurs statistiques qui permettrait d'expliquer la variation de l'indice boursier. Ce type de technique suppose en effet que l'information extérieure est automatiquement incorporée dans les prix (efficience semi-forte) et donc l'analyse de ces indicateurs devient inutile. Cependant, elle fait l'hypothèse que les prix récents peuvent influencer les cours actuels. Le tableau suivant résume les principales caractéristiques techniques statistiques utilisées :

Indicateurs	Formule
MA_n	$\sum_{i=1}^{n} P_{t-i}$
stochastic $K_n\%$	$100\frac{P_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}}$
stochastic $D_M\%$	$\frac{1}{M}\sum_{i=1}^{M} K_i$
slow D_M	$\frac{1}{M}\sum_{i=1}^{M} D_i$
Momentum_n	$P_t - P_{t-n}$
ROC_n	$100\frac{P_t}{P_{t-n}}$
William's R%	$100\frac{HH_{t-n} - P_t}{HH_{t-n} - LL_{t-n}}$
A/D Oscollator	$\frac{H_t - P_{t-1}}{H_t - L_t}$
OSCP	$\frac{MA_5 - MA_{10}}{MA_5}$

où P_t représente le prix de l'indice au temps t, LL_t et HH_t respectivement les plus petits et plus grands prix durant les t dernières périodes.

La seconde approche fondamentale s'appuie sur des facteurs extérieurs pouvant fluctuer le cours de l'indice (indicateurs économiques, analyse de sentiment, actualité politique ...). Cette méthode repose sur l'étude des facteurs pouvant influencer l'évolution de l'offre et de la demande. Dans notre cas, nous utilisons des indicateurs macroéconomiques pour prédire la direction du cours de Nikkei 225. Plus particulèrement, l'indice SP&500 et la parité USD/JPY sont considérer comme des variables influentes puisque les importants échanges commerciaux entre le Japon et l'Amérique marquent une interdépendance de ces marchés. D'autres facteurs fondamentales sont utilisés dans la littérature tels que les taux d'intérêts à court et long terme, l'indice des prix à la consommation, la production industrielle, la dépense publique, la consommation privée, le PNB ou le PIB.

II - Revue des méthodes d'apprentissage employées

1 - Classifieurs binaires

• LDA et QDA

On se propose dans cette section de décrire les algorithmes de classification Linear Discriminant Analysis (LDA) et Quadratic Linear Analysis (QDA). Ces deux algorithmes étant quasi similaires, nous déduirons la frontière de décision du classifieur QDA à partir de celle de la LDA.

L'approche principale du formalisme de LDA repose sur la formule de Bayes :

$$\mathbb{P}(Y = i|X \in A) = \frac{\mathbb{P}(X \in A|Y=i)\mathbb{P}(Y=i)}{\mathbb{P}(X \in A|Y=-1)\mathbb{P}(Y=-1)+\mathbb{P}(X \in A|Y=1)\mathbb{P}(Y=1)}$$

avec i∈{-1,1} (Y=1 si direction à la hausse, -1 si à la baisse) et A∈ $\mathbb{R}^d$ (d=2 : SP&500 et USD/JPY)

On suppose à présent que la loi conditionnelle de $X|Y = i$ est absolument continu par rapport à la mesure de Lebesgue dans $\mathbb{R}^2$ on a alors $\mathbb{P}_{X|Y=i}(dx) = f_i(x)dx$ avec f_i la densité de la loi conditionelle. On a vu dans le Notebook Jupiter que cette loi conditionnelle suivait des lois gaussiennes après transformation logarithmique des prix. Ainsi, on suppose que f_i est la densité d'une loi gaussienne de paramètre (μ_i, Σ) avec la même matrice de covariance Σ pour les 2 lois conditionnelles. La formule se réécrit ainsi :

$$\mathbb{P}(Y = i|X \in A) = \frac{f_i(x)\mathbb{P}(Y=-1)}{f_{-1}(x)\mathbb{P}(Y=-1)+f_1(x)\mathbb{P}(Y=1)}$$

Et alors la classification naturellement $\widehat{f}$ est donnée par :
$\widehat{f}$(x)=$\begin{cases} 1 \text{ si } \mathbb{P}(Y=1) > \mathbb{P}(Y=-1) \\ \text{-1 sinon} \end{cases}$
Et on trouve après calcul la frontière de décision suivante :

$$D(x) \geq 1 \iff \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=-1)} \geq 1 \iff x^t\Sigma^{-1}(\mu_1 - \mu_{-1}) \geq \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1 - \mu_{-1}^t\Sigma^{-1}\mu_{-1}$$

La frontière de décision est donc linéaire (de la forme $x^t a + b \geq 0$). Il reste à déterminer les paramètres (μ_i, Σ) qui sont obtenus par estimation empirique de Monte Carlo.

À la différence de la LDA, la QDA suppose que chacune des 2 lois conditionnelles ont des matrices de covariances différentes notées Σ_i Par analogie, on trouve une frontière de décision quadratique. En pratique, la méthode QDA peut sembler plus judicieuse que la LDA puisqu'elle n'impose aucune contrainte sur les matrices de covariances conditionnelles. Cependant, dans le cas où la taille de l'échantillon de donnée est insuffisant, les estimateurs de Monte Carlo seront imprécis et la QDA aura alors un nombre plus élevé d'estimateurs biaisés comparé à la LDA, engendrant un modèle complexe faussé.

• SVM

Le classifieur SVM (Support vector Machine) introduit par Vladimir Vapnik en 1995 est largement utilisé dans la prédiction des séries financières et s'avère être très souvent le plus performant. Par exemple, Kim (2003) examine la faisabilité de l'application de SVM dans la prévision financière en la comparant à des réseaux neuronaux de rétropropagation. L'analyse des résultats expérimentaux a montré qu'il est avantageux d'appliquer des SVM pour prévoir les séries chro-

nologiques financières. L'article que l'on étudie en arrive à la même conclusion.

Introduisons tout d'abord un fonction $\Phi : \mathbb{R}^d \to \mathbb{R}^M$ qui transforme nos variables explicatives initiales (d=2 ici) dans un espace à plus grande dimension M tel que les donnée soient séparable linéairement. Ainsi, $\exists (u, v) \in (\mathbb{R}^M)^2$ tel que $\{x, u^t\Phi(x) + v = 0\}$ représente l'hyperplan séparant nos données en 2 classes distinctes. Quitte à interchanger les étiquettes Y de nos observations, on suppose que $\{x_i, u^t\Phi(x_i) + v > 0\}_{i\leq n}$ sont dans la classe {Y=1} et $\{x_i, u^t\Phi(x_i) + v < 0\}_{i\leq n}$ sont dans la classe {Y=-1}. Cependant le couple séparateur (u, v) n'est jamais unique. On va chercher l'unique hyperplan séparateur à vaste marge qui est tel que la plus petite distance entre nos observations $(\Phi(x_i))_{i\leq n}$ et l'hyperplan $\{x, u^t\Phi(x) + v = 0\}$ soient supérieur à un réel positif K. Puis on maxime cette distance K. On obtiendra donc un hyperplan séparant nos 2 classes de façon optimale. En prenant la distance géométrique d'un point x par rapport à l'hyperplan ($d(x, (u, v))=\frac{|u^t\Phi(x)+v|}{\|u\|}$), le problème d'optimisation s'écrit :

$$\begin{cases} \max\limits_{(u,v)} \text{K} \\ \text{tq } \min\limits_{i\leq n} \frac{|u^t\Phi(x_i)+v|}{\|u\|} \geq \text{K} \end{cases}$$

En posant $a = \frac{u}{\|u\|K}$ et $b = \frac{v}{\|u\|K}$, on obtient $K = \frac{1}{\|a\|}$. Or maximiser $\frac{1}{\|a\|}$ est équivalent à minimiser $\|a\|^2$, et en utilisant les étiquettes y_i on trouve le problème équivalent à une constante multiplicative près :

$$\begin{cases} \min\limits_{(a,b)} \frac{\|a\|^2}{2} \\ y_i(a^t\Phi(x_i) + b) \geq 1 \ \forall i \leq n \end{cases}$$

Avant de résoudre ce système, on va modifier notre hypothèse initiale qui suppose que nos données peuvent être linéairement séparable dans un espace à plus grande dimension. En pratique ce n'est quasiment jamais possible à cause de certains points anormales possédant des caractéristiques types d'une autre classe. Afin d'alléger nos hypothèses, on assume que certains points peuvent être mal classés (par exemple $a^t\Phi(x_i) + b < -1$ alors que $y_i = 1$: le point i se trouve du mauvais côté de la frontière) ou seulement à une plus petite distance que la marge K fixée ($0 \leq a^t\Phi(x_i) + b \leq 1$ et $y_i = 1$), en introduisant les points positifs $(\epsilon_i)_{i\leq n}$ rendant la distance des points à la frontière flexible :

$$\begin{cases} \min\limits_{(a,b)} \frac{\|a\|^2}{2} + C\sum\limits_{i\leq n} \epsilon_i \\ y_i(a^t\Phi(x_i) + b) \geq 1\text{-}\epsilon_i \ \forall i \leq n \end{cases}$$

Dès que ϵ_i dépasse 1, on autorise le point x_i à être de l'autre côté de la frontière (mal classé). Cependant, on rajoute une contrainte sur les $(\epsilon_i)_{i\leq n}$ afin de pénaliser les points mal classés ou inférieur à la marge. Cette pénalisation est contrôlée par la constante C>0 : plus la constante est grande plus les points $(\epsilon_i)_{i\leq n}$ diminuerons et donc moins de points seront autorisés à s'affranchir des contraintes de la marge.

On trouve ensuite l'unique solution du problème d'optimisation avec le théorème de Kuhn-Tucker et on obtient :

$$a^* = \sum_{i\leq n} \alpha_i^* y_i \Phi(x_i)$$

$$b^* = \frac{1}{N} \sum_{i,0<\alpha_i^*<C} (y_i - a^{*t}\Phi(x_i))$$

avec $(\alpha_i^*)_{i\leq n}$ les multiplicateurs solutions du Lagrangien.

La frontière de décision D(x) est donc l'hyperplan optimal définit par (a^*, b^*) :

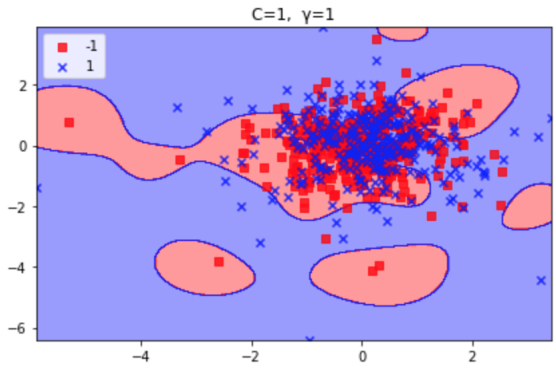
$$D(x) = \{x, (\sum_{i\leq n} \alpha_i^* y_i \Phi(x_i)^t \Phi(x) + \frac{1}{N} \sum_{j,0<\alpha_j^*<C} (y_j - \sum_{i\leq n} \alpha_i^* y_i \Phi(x_i)^t \Phi(x_j))) = 0\}$$

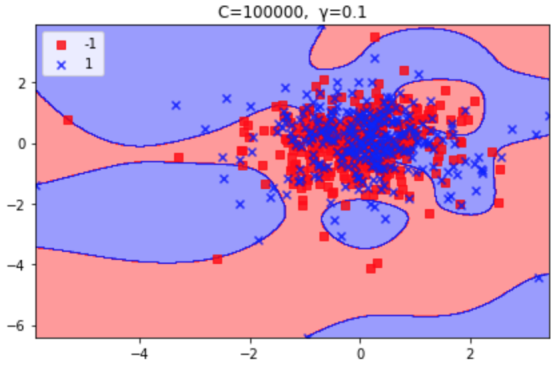
On remarque que l'on doit calculer un produit scalaire entre les vecteurs de dimension M $\Phi(x_i)$ et $\Phi(x_j)$ ce qui s'avère très couteux en temps de calcul. On introduit donc des fonctions noyau notées k qui égalisent ce produit scalaire sous les hypothèses du théorème de Mercer : $k(x_i, x_j) = \Phi(x_i)^t \Phi(x_j)$. Le noyau gaussien ($k(x, y) = e^{-\gamma\|x-y\|^2}, \gamma > 0$) vérifie ces hypothèses et est généralement utilisé dans la prédiction des séries financières qui tend à donner de bons résultats sans hypothèse particulière sur les données.

Finalement, notre classifieur SVM est : $\widehat{f}(\text{x}) = \begin{cases} 1 \text{ si D(x)} > 0 \\ \text{-1 sinon} \end{cases}$

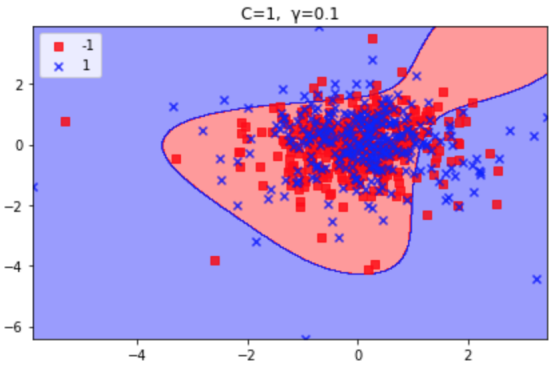
Ainsi, les 2 paramètres C et γ restent à determiner par méthode de Cross Validation. On montre ci dessous l'influence des paramètres sur la frontière de décision entrainée sur les données de l'article étudié.

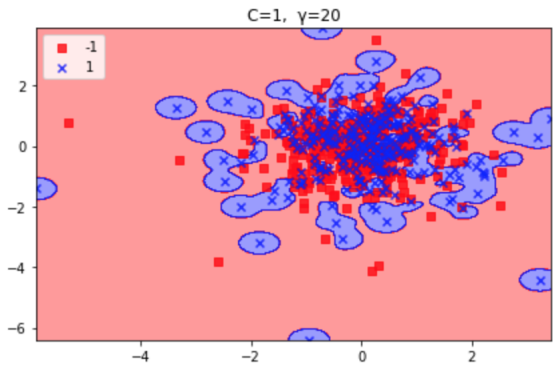
On remarque que la paramètre C tend à contrôler les erreurs de classification : à γ fixé, plus C augmente moins il y a d'erreur impliquant un risque de sur apprentissage :





Le paramètre γ contrôle la zone d'influence des points vis à vis de la frontière de décision : plus γ augmente plus la frontière de décision tend à se rapprocher individuellement de chacun des points impliquant un risque de sur apprentissage sévère. Si γ devient trop large, la valeur de C n'aura plus aucune influence et ne pourra donc plus contrôler le taux de classification erroné.





2 - Réduction de dimension

Dans notre Notebook Jupiter, nous avons tenter d'améliorer la robustesse du classifieur SVM en appliquant de nouvelles caractéristiques fondamentales (13 au total). Afin de garder les variables les plus influentes et de diminuer les temps de calcul, nous avons introduit 2 méthodes de réduction de dimension : le Test de Fisher et l'ACP (Analyse en Composantes Principales)

- Test de Fisher

Le test de Fisher calcul un score pour chacune des variables explicatives. Un haut score implique une plus grande capacité à prédire la variable à expliquer Y et inversement. Il suffira alors de garder les K variables avec le plus haut score pour obtenir une réduction à K dimension. Le scoring de la i-ème variable est défini par :

$$S_i = \frac{\sum_{j \in \{-1,1\}} n_j (\mu_{i,j} - \mu_i)^2}{\sum_{j \in \{-1,1\}} n_j \sigma_{i,j}^2}$$

avec n_j le nombre d'observations dans la classe j, $\mu_{i,j}$ et $\sigma_{i,j}^2$ respectivement la moyenne et la variance des observations de la variable i qui sont dans la classe j.

Plus précisément on fixe une caractéristique i, on calcule la moyenne de toutes ses observations que l'on compare aux moyennes inter-classes (les variables i étiquetées -1 et celles étiquetées 1). Si les points de cette variable sont bien séparés par classe, on obtiendra 2 groupes distincts situés de part et d'autre de la moyenne de toutes les observations et le

numérateur sera non nulle. Inversement si les points sont mal classés, les moyennes inter-classes risquent de se confondre à la moyenne total et le numérateur sera proche de 0. On normalise ensuite ce score avec les variances inter-classes pour pénaliser les variables trop dispersées. Cette méthode a l'avantage d'être interprétable mais reste simpliste dans son formalisme et ne permet pas de quantifier d'éventuelle lien entre les différentes caractéristiques. Par exemple, il se peut que deux variables prisent individuellement aient un score faible mais qu'une combinaison de ces attributs améliore fortement ce score.

- ACP

La réduction de dimension par l'ACP repose sur une minimisation de l'inertie mesurant la dispersion des points. L'ACP projette le nuage de point initial d'un espace à grande dimension dans un nouveau espace à plus petite dimension de façon à perdre le minimum d'information (dispersion) des points.
Le nuage de point est dans un premier temps centré (chaque variable se voit retirer sa moyenne). Puis on cherche un axe réel noté δ de vecteur directeur u unitaire tel que l'inertie du nuage projeté sur cette axe soit maximale (on passe d'un nuage de n vecteurs de dimension p à n points réels). L'inertie est la somme des carrées de ces nouveaux points. Plus formellement, en notant C le vecteur de taille n des points de l'axe δ on a : $C = X.u$ d'inertie $C^tC = u^tX^tXu$. L'inertie de C correspond ici à sa variance empirique puisque X est centré implique C centré. On veut donc résoudre :

$$(*) \max_{u^tu=1} u^tX^tXu$$

On trouve facilement avec le théorème de KKT et en notant λ le mul-

tiplicateur positif du Lagrangien :

$$X^t X u = \lambda u$$
$$u^t u = 1$$

On remarque donc que u est le vecteur propre normé associé à la valeur propre λ de la matrice $X^t X$ qui est symétrique réel positive et donc toujours diagonalisable dans une base orthogonale avec des valeurs propres positives. La maximisation en u de (*) revient donc à maximiser sa valeur propre λ : u est le vecteur propre normé associé à la plus grande valeur propre de $X^t X$. D'autre part on a également que la variance empirique de C est $\frac{1}{n} C^t C = \frac{1}{n}\lambda$ et la somme des valeurs propres de $X^t X$ est la trace de cette matrice qui est également la trace de la matrice de variance empirique de X à une constante multiplicative près. Ainsi, $\frac{\lambda_{max}}{\sum \lambda_i}$ représente la quantité de variance expliquée par l'axe δ parmi la variance totale ou encore la proportion d'information captée par δ.

En réitérant ce procédé sur un nouvel axe ω de vecteur directeur unitaire v orthogonal à u, on obtient que v est le vecteur propre de $X^t X$ associé à la seconde plus grande valeur propre (car la matrice de passage est orthogonale ie que les vecteurs propres sont mutuellement orthogonaux). Cette axe captera donc forcément moins d'information que le précédent. On impose à ce que v soit orthogonal à u afin d'obtenir de nouvelles informations non expliquées par l'axe δ.

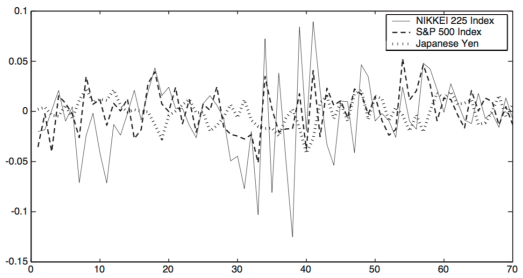
On continue ce raisonnement jusqu'à obtenir un certain pourcentage cumulé de variance expliquée (typiquement 90%). L'ACP a donc l'avantage de toujours garder les p variables explicatives initiales même en réduisant la dimension sur la droite réelle puisque la projection des données sur un axe se réalise par un produit scalaire entre les points individus et le vecteur directeur, c'est-à-dire par une combinaison linéaire

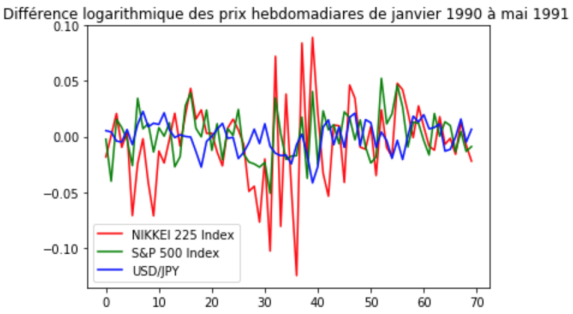
des variables explicatives.

I - Analyse des résultats

1 - Reproduction des résultats de l'article

Dans la première partie du Notebook Jupiter, on a tenté de retrouver les résultats de l'article. On a importé les mêmes données mentionnées par l'article via le moteur de recherche "Yahoo finance" (pour les prix de l'indice Nikkei 225 et de l'indice SP&500) et via le site du professeur Werner Antweiler qui héberge les données de la parité USD-JPY. On s'est assuré que les données de l'article concordent bien avec les nôtres :





Puis on a appliqué une Cross Validation de type Shuflle Stratifié pour déterminer les paramètres (C, γ) du SVM. Cette Cross validation consiste à mélanger initialement les données à entrainer puis les séparer en 2 (80% contre 20%) de sorte que la séparation induisent la même proportion initiale des classes représentées dans les 2 sets séparés. Cette opération de séparation stratifiée sera répétée 5 fois. Cette méthode permet d'obtenir un biais moins important sur la performance finale. Cependant, cette méthode renvoyée un couple de pa-

ramètre très différents entre plusieurs Cross Validation de même type. Afin de stabiliser les paramètres optimaux et d'obtenir une unique solution, on a appliqué la Leave One Out Cross Validation (LOOCV). Cette méthode repose sur un entrainement de toutes les valeurs possibles du train set auquel on a retiré un unique point qui sera testé. On réitère cette opération jusqu'à que tous les points du train set aient été enlevés une unique fois. Ainsi avec cette méthode la Cross Validation ne pourra renvoyer qu'un unique couple quel que soit le nombre de Cross Validation lancée (il n'y a plus d'aléa).

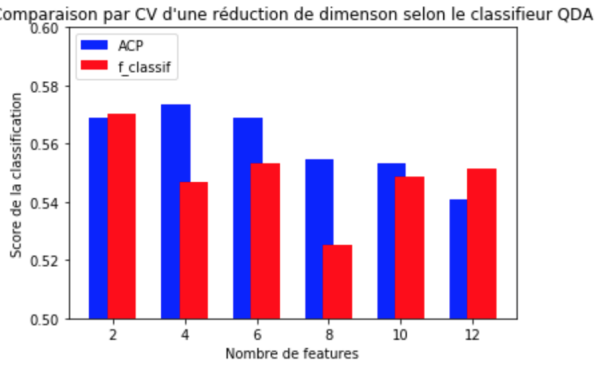
L'unique couple de paramètre du SVM renvoyé par la LOOCV donne un hit ratio de 66% contre 72% pour l'article. Avec la LDA et la QDA on a obtenu un score de respectivement 55% et 0.61% contre respectivement 0.61% et 0.69% dans l'article. Cet écart peut semblé important mais ramené à la faible quantité de données test (36), nos classifieurs ont prédisent seulement 2 à 3 points en moins en comparaison de l'article.

Cette observation nous a naturellement amenée à élargir le test set en ajoutant de nouvelles données de tel sorte d'obtenir 20% de données test contre 5% auparavant. L'excellente performance des classifieurs chute drastiquement suite à ce changement : la LDA diminue de 5% tandis que la QDA et le SVM perdent plus de 10 points ! D'autre part, une analyse plus poussée de ces résultats nous a amenée à la conclusion que le SVM était très performant sur les premies 36 points test suivant la période d'entrainement mais bien moins compétitif pour les points suivants plus éloignés. Il semble que la capacité de prédiction du SVM décroit avec le temps.

2 - Amélioration de la stabilité de la performance

Afin de pallier à la chute de la performance au cours du temps du classifieur SVM, on a décidé d'implémenter notre propre modèle de prédiction du Nikkei 225. Au total, 13 variables macroéconomiques ont été importées (indice boursier américian, européen et chinois ; taux intérêt des obligations d'États à maturité 1, 2, 3, 5, 7, 10, 15 - parité USD-JPY, EUR-JPY, CNY-JPY) auxquelles on a racolé les valeurs limites au borne d'un intervalle de validation puis les variables ont été normalisées dans [-1,1]. Cette procédure permet d'augmenter la performance des classifieurs (cf Francis E.H. Tay, L.J. Cao (2001)) et de diminuer les temps de calculs.

Nous avons implémenté notre algorithme de Cross Validation afin de pouvoir appliquer les 2 méthodes de réduction présentées dans la section précédente. L'ACP donne pour presque toutes les dimensions réduites testées de meilleurs performance pour un classifieur fixé, comparé au Test de Fisher. Ci-dessous les résultats obtenus par la QDA :



Nous choisissons donc d'utiliser l'ACP comme méthode de réduction pour nos classifieurs. La dimension réduite est fixée à 8 de sorte à obtenir 90% de variance expliquée. Le test de Fisher permet également de nous rendre compte des variables individuelles ayant le plus d'influence sur la direction du Nikkei 225 comme le montre la capture suivante :

```
Feature selection par 'f_classif' :

    2 meilleures features : ['YB5', 'YB10']
    4 meilleures features : ['YB3', 'YB5', 'YB7', 'YB10']
    6 meilleures features : ['YB2', 'YB3', 'YB5', 'YB7', 'YB10', 'STOXX50E']
    8 meilleures features : ['YB2', 'YB3', 'YB5', 'YB7', 'YB10', 'YB15', 'STOXX50E', 'USDJPYbis']
    10 meilleures features : ['YB1', 'YB2', 'YB3', 'YB5', 'YB7', 'YB10', 'YB15', 'STOXX50E', 'SSE', 'USDJPYbis']
    12 meilleures features : ['YB1', 'YB2', 'YB3', 'YB5', 'YB7', 'YB10', 'YB15', 'SP500bis', 'STOXX50E', 'SSE', 'USDJ
PYbis', 'EURJPY']
```

On constate que les taux d'intérêts des obligations d'États (notés YB) ont un fort impact sur la variable à expliquer. Cela peut s'expliquer par le fait que ces taux d'intérêts redent compte de l'inflation du Japon et donc de la consommation des ménages. En effet, les taux à courts termes fixés par la BoJ permettent principalement de contrôler le taux d'inflation du pays : en cas de hausse des prix, les taux d'intérêts augmentent impliquant une hausse des taux d'intérêts des crédits bancaires et donc une baisse de la consommation des ménages. Les investisseurs demanderont en parallèle des taux de coupons plus élevés pour effacer les effets de l'inflation impliquant une baisse automatique des prix des obligations. Ces taux d'intérêts reflètent donc bien l'économie générale du pays.

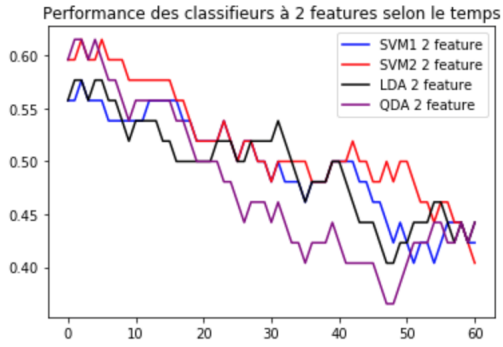
Après avoir obtenus le couple de paramètre avec la LOOCV, on obtient les résultats du SVM avec ACP. Les matrices de confusions pour 20% de points test du SVM à 2 variables et du SVM à 13 variables

avec ACP sont respectivement comparées ci dessous :

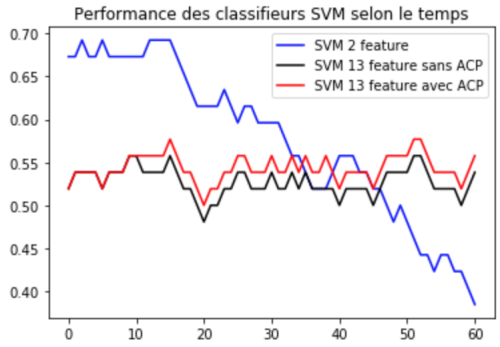
	precision	recall	f1-score	support
Baisse	0.48	0.78	0.59	49
Hausse	0.68	0.36	0.47	64
avg / total	0.59	0.54	0.52	113

	precision	recall	f1-score	support
Baisse	0.46	0.89	0.60	46
Hausse	0.78	0.27	0.40	67
avg / total	0.65	0.52	0.48	113

En s'intéressant à la précision et au rappel, on remarque que le SVM avec ACP obtient une précision supérieur pour un rappel légèrement inférieur au SVM initial à 2 caractéristiques. Notre nouveau SVM est donc plus précis que l'ancien ce qui peut s'expliquer par une indépendance temporelle. Pour vérifier cela, nous décidons de tester nos classifieurs sur différentes périodes test de 52 points (52semaines = 1an) chacune décalées d'1 jour. On vérifie d'abord la propriété de dépendance temporelle des anciens classifieurs QDA, LDA et SVM :



Puis on montre l'indépendance de nos nouveaux classifieurs SVM :



Le graphique ci dessus est sans equivoque : le simple SVM à 2 caractéristiques de l'article est fortement dépendant du temps. Seuls les points test les plus proches temporellement de notre période d'entrainement sont correctement classifiés. Avec le temps, le classifieur devient incompétent et ses anciennes données ne sont plus capable de prédire de nouveaux points tests. Par exemple, 15 semaines après la fin de l'entrainment, la performance sur 1 an chute drastiquement de plus de 10%. En revanche, nos nouveaux classifieurs mis en place à 13 caractéristiques semblent indépendant de la variable temporelle puisque qu'elle que soit la période test la performance du hit ratio varie toujours entre 51 et 57%.

D'autre part, on voit également la sur-performance à tout instant de l'application de l'ACP sur les 13 features (la courbe rouge est toujours au dessus de la courbe noire). Cette observation met en exergue le principe de parcimonie : apporter de nouvelles informations en plus à un jeu de donnée n'implique pas forcément une meilleure classification.

En conclusion, les précédents graphiques montrent qu'il est nécéssaire que les classifieurs à 2 caractéristiques assignent correctement les premiers points test suivant directement la fin de la période d'entrainement. Cependant, cette tâche s'avère compliquée puisqu'il faut obtenir un couple de paramètre par Cross Validation prédisant le mieux les points test récents. Il aurait sans doute était préférable d'introduire une notion de temporelle dans la Cross Validation en découpant par exemple les données à entrainer en un train set des points les plus anciens pour la prédiction d'un test set suivant temporellement le train set. Puis on réitère cette opération en décalant de 1 point notre train set et test set jusqu'à que le dernier point du test set soit le dernier point de l'ensemble d'entrainement initial. Sans un tel processus dépendant du temps, les paramètres ne donneront pas obligatoirement de bonne performance sur les points à tester les plus récents et la performance sera moindre.

Nous avons pu contourner cette étape en augmentant le nombre de caractéristiques engendrant une indépendance temporelle pour une performance moindre mais stable (de l'ordre de 54.5% en moyenne variant de 52% à 57%) et une précision supérieure.

Enfin, un article de recherche datant de 2001 ("Modidied support vector machines in dinancial time series forecasting", Francis E.H. Tay, L.J. Cao) propose de modifier le problème d'optimisation du SVM afin d'apporter une plus grande importance aux données à tester les plus récentes. La méthode est appliquée sur la prédiction des prix du SP&500 (régression) et il aurait pu être intéressant d'appliquer ces résultats pour une classification binaire de la direction du Nikkei 225.

Sources

"Predicting direction of stock price index movement using artificial neural networks and support vector machines : The sample of the Istanbul Stock Exchange", Yakup Kara, 2011

"Modifed support vector machines in nancial time series forecasting", Francis E.H. Tay, L.J. Cao, 2001

"Improved financial time series forecasting by combining Support Vector Machines with self-organizing feature map", Francis Eng Hock Tay and Li Juan Cao, 2001

"Application of support vector machines in nancial time series forecasting", Francis E.H. Tay, Lijuan Cao, 2001

"Support vector machines experts for time series forecasting", Lijuan Cao, 2002

"Introduction to financial forecasting", YaserAbu-Mostafa, 1996

"Support Vector Machines approach to predict the S&P CNX Index returns", Mr. Manish Kumar & Dr. M. Thenmozhi, 2007

"Forecasting stock index movement : a comparaison of Support Vector Machines and Random Forest", Manish Kumar & M. Thenmozhi, 2007

"Business conditions and expected returns on stocks and bonds", Fama E, French K, 1990