MACHINE LEARNING

1st Assignment

ΠΑΡΑΣΚΕΥΟΠΟΥΛΟΥ ΓΕΩΡΓΙΑ

Πρόβλημα 1.1

Τα δεδομένα σε αυτό το πρόβλημα, καθώς και σε όλα τα υποερωτήματα του προβλήματος 1 τα δεδομένα κατασκευάστηκαν ακολουθλωντας το μη γραμμικό μοντέλο :

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 {x_n}^2 + \theta_3 {x_n}^3 + \theta_4 {x_n}^4 + \theta_5 {x_n}^5 + \eta_n,\ n = 1,2,\dots,N.$$

Για την εκτίμηση των συντελεστών του πολυωνύμου χρησιμοποιήθηκε η Least Square Method σύμφωνα με την οποία :

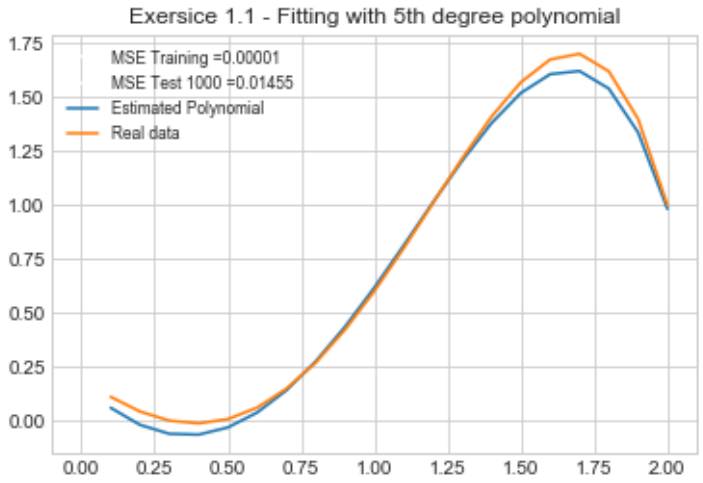
$\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi \boldsymbol{y}$, όπου

$\Phi = [1, \varphi_1(\boldsymbol{x}) \quad \varphi_2(\boldsymbol{x}) \dots \quad \varphi_{K-1}(\boldsymbol{x})]^T$

και $\varphi_i(x) = x^i, i = 1,..,K-1.$

Εδώ K=6, εφόσον το πολυώνυμο παλινδρόμησης ήταν 5ου βαθμού και $\theta = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5]$.

Στο σημείο αυτό να σημειωθεί πως σε όλα τα ερωτήματα του προβλήματος 1 ακολουθήθηκε αυτός ο τρόπος κατασκευής του πίνακα Φ, αλλάζοντας την τιμή του K σύμφωνα με τον βαθμό του πολυωνύμου παλινδρόμησης.

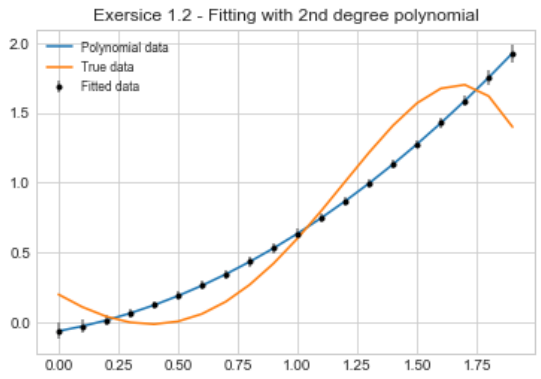


Γράφημα 1.1.

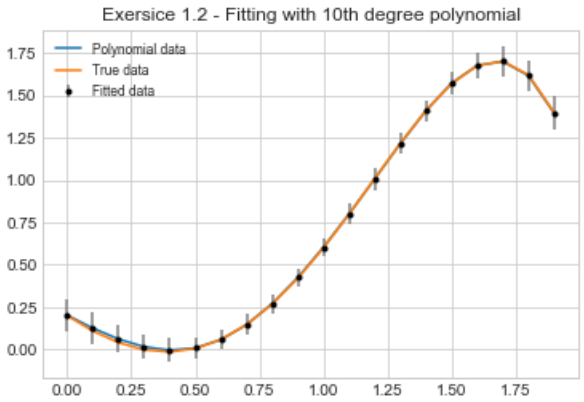
Όπως φαίνεται τώρα στο γράφημα το MSE για το training set είναι πολύ μικρό $MSE_{training}$=0.00001 που σημαίνει ότι η εκτίμηση (fitting) είναι πολύ κοντα στα training points. Επίσης, $MSE_{testing}$=0.01455, οπού και αυτό είναι αρκετά μικρό αλλά αρκετά μεγαλύτερο από το $MSE_{training}$, κάτι το οποίο είναι λογικό γιατί το μοντέλο εκπαιδεύτηκε με το training set, ενώ το test set αποτελείται από 1000 διαφορετικά σημεία με noise διασποράς ${\sigma_\eta}^2 = 0.1$.

Προβλημα 1.2

Στο πρόβλημα αυτό χρησιμοποιήθηκε και πάλι η μέθοδος Least Square για την εύρεση της συνάρτησης παλινδρόμησης με την εξής όμως διαφορά , το πολυώνυμο που χρησιμοποιήθηκε ήταν τη μια φορά 2ου βαθμού και 10ου βαθμού την άλλη. Στα παρακάτω γραφήματα φαίνονται : α) το πραγματικό πολυώνυμο (5ου βαθμού) από το οποίο κατασκευάστηκε το training set β) οι μέσοι όροι των εκτιμήσεων του y και οι αντίστοιχες διακυμάνσεις για τα 20 training points όπως αυτές προέκυψαν μετά από 100 επαναλήψεις του πειράματος με διαφορετικό noise στο training set κάθε φορά.



Γράφημα 1.2.1.



Γράφημα 1.2.2.

Τα αποτελέσματα που προέκυψαν ήταν τα αναμενόμενα και επιβεβαιώνουν το Bias-Variance Dilemma. Συγκεκριμένα, στην περίπτωση που το fitting γίνεται με πολυώνυμο 2^{ου} βαθμού (Γράφημα 1.2.1.), δηλαδή α αριθμός των παραμέτρων προς εκτίμηση είναι μικρός παρατηρούμε ότι το variance των εκτιμήσεων του y ανάμεσα στα 100 πειράματα είναι πάρα πολύ μικρό, σε ανίθεση με το bias που είναι εξαιρετικά μεγάλο, δεδομένου ότι τα εκτιμώμενα y διαφέρουν πάρα πολύ από τα πραγματικά. Από την άλλη, στην περίπτωση που το πολυώνυμο είναι 10^{ου} βαθμού (Γράφημα 1.2.2), δηλαδή ο αριθμός των παραμέτρων προς εκτίμηση είναι μεγάλος παρατηρούμε ότι το variance είναι μεγάλο σε σύγκριση με την περίπτωση του πολυωνύμου 2^{ου} βαθμού. Αντίθετα, το bias είναι πάρα πολύ μικρό και οι εκτιμήσεις του y είναι πολύ κοντά στις πραγματικές τιμές.

Προβλημα 1.3

Εδώ επαναλήφθηκε το πείραμα του Προβλήματος 1 χρησημοποιώντας πολυώνυμο 5^{ου} βαθμού, όμως η παλινδρόμηση έγινε τη μέθοδο Ridge Regression σύμφωνα με την οποία :

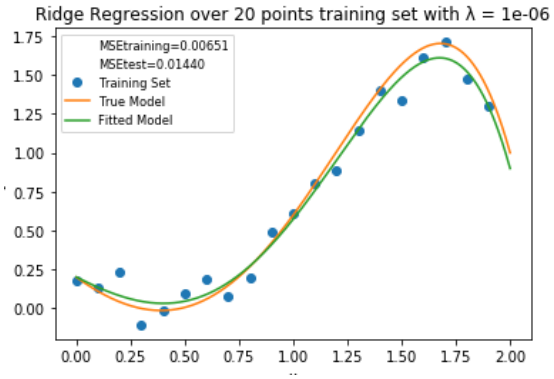
$$\hat{\theta} = (\Phi^T \Phi - \lambda I)^{-1} \Phi \boldsymbol{y} \ .$$

Δοκιμάστηκαν διάφορες τιμές του λ ([0.000001, 0, 0.001, 0.003, 0.0035, 0.0025, 0.0055, 0.1, 0.3 , 1.0, 10.0, 100.0, 1000]) και στα επόμενα γραφήματα φαίνονται τα αποτελέσματα για κάποιες ενδεικτικές. Για το σχολιασμό των αποτελεσμάτων και τη σύγκριση ανάμεσα στις διάφορες τιμές του λ χρησιμοποιήθηκε και η τιμή λ=0 , σύμφωνα με την οποία προκύπτουν τα αποτελέσματα της παλινδρόμησης σύμφωνα με τη μέθοδο Least Square.

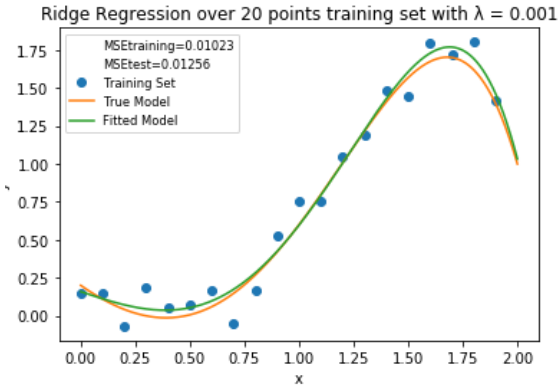
Παρατηρούμε λοιπόν πως όταν το λ γίνεται πολυ μικρό (λ=1e-06, Γράφημα 1.3.2) τα αποτελέσματα μοιάζουν πολύ με αυτά της Least Square (Γράφημα 1.3.1), όπως ήταν αναμενόμενο. Καθώς όμως το λ πλησιάζει να γίνει 0.03 (εδώ έχουμε παρουσιάσει μόνο την περίπτωση λ=0.03, Γράφημα 1.3.4) $MSE_{testing}$ βελτιώνεται, και όταν δωθούν τιμές μεγαλύτερες του 0.03, ακόμα και η τιμή λ=0.035 το $MSE_{testing}$ αρχίζει να αυξάνεται.Στο σημείο αυτό αξίζει να σχολιαστούν τα αποτελέσματα για μεγάλες τιμές του λ (ενδεικτικά παραουσιάζεται στο Γράφημα 1.3.6 η περίπτωση λ=1000). Παρατηρήθηκε λοιπόν οτι καθώς το λ αυξάνεται πολύ δηλαδή αυξάνεται πολύ και το bias που βάζουμε στον εκτιμητή τα MSE αυξάνονται

επίσης και μάλιστα σημαντικά, όπως ήταν αναμενόμενο, αφού καθώς το λ αυξάνεται εξαλείφονται οι κορυφές της καμπύλης παλλινδρόμησης κάνοντας την πιο smooth.

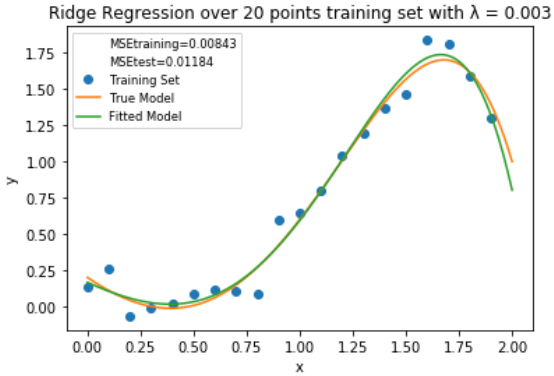
Γράφημα 1.3.1.



Γράφημα 1.3.2.

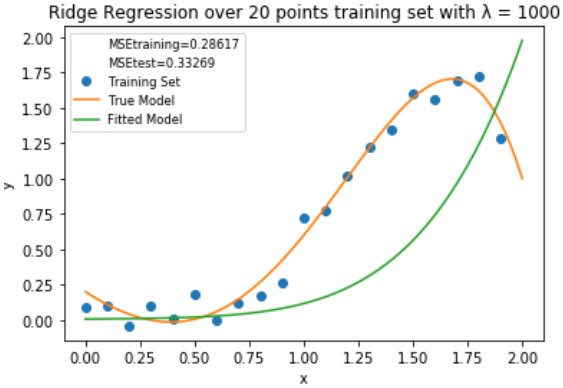


Γράφημα 1.3.3.



Γράφημα 1.3.4.

Γράφημα 1.3.5.

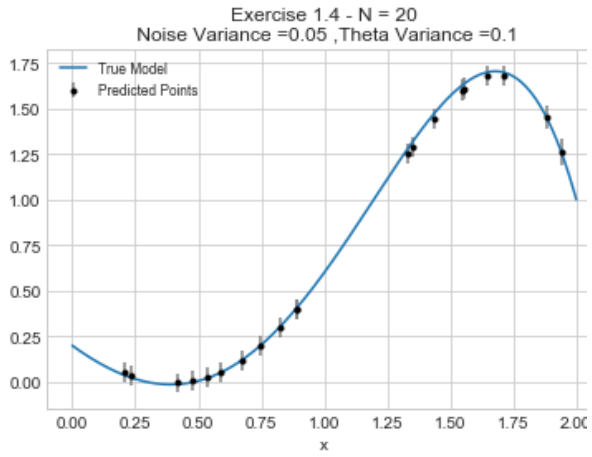


Γράφημα 1.3.6.

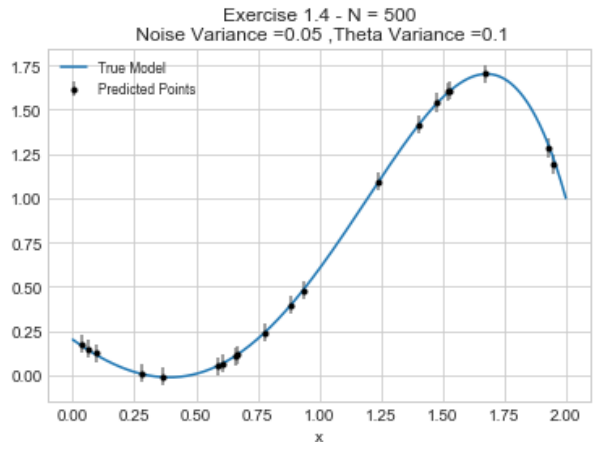
Προβλημα 1.4

Στο πρόβλημα αυτό χρησιμοποιήθηκε το σωστό μοντέλο και ως μέση τιμή $\boldsymbol{\theta_0}$ της prior κατανομής του $\boldsymbol{\theta}$ το πραγματικό $\boldsymbol{\theta}$ από το οποίο κατασκευάστηκε το training set. Στα γραφήματα 1.4.1-1.4.4. φαίνονται τα αποτελέσματα της εκτίμησης των respective $\hat{y}$ (μ_y) για τις αντίστοιχες τιμές των 20 τυχαίων $x \in [0,2]$,

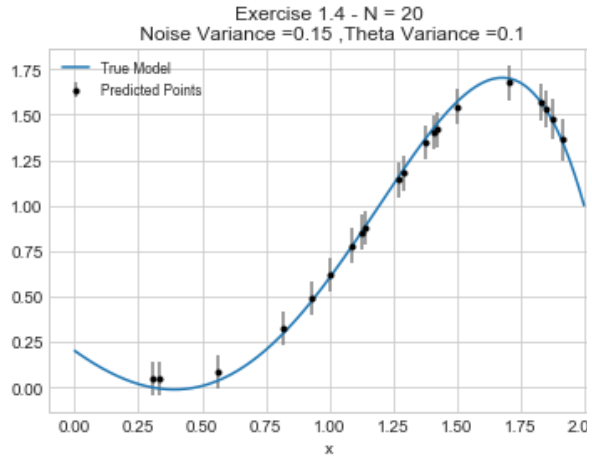
σύμφωνα με τη Bayesian προσέγγιση. Επιπλέον με γκρι απεικονίζονται τα error bars για κάθε εκτίμηση του y, τα οποία υπολογίστηκαν βάσει των αντίστοιχων διακυμάνσεων $\sigma_y{}^2$. Δοκιμάστηκαν δυο τιμές για το πλήθος των σημείων του training set, N=20 και N=500. Γενικά, παρατηρούμε ότι καθώς αυξάνεται ο αριθμός των training points οι εκτιμήσεις βελτιώνονται και πως καθώς το αυξάνεται το $\sigma_\eta{}^2$ μεγαλώνουν και τα error bars.



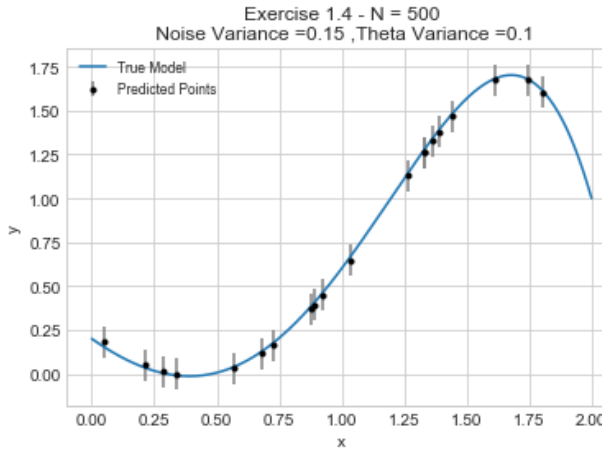
Γράφημα 1.4.1.



Γράφημα 1.4.2.



Γράφημα 1.4.3.

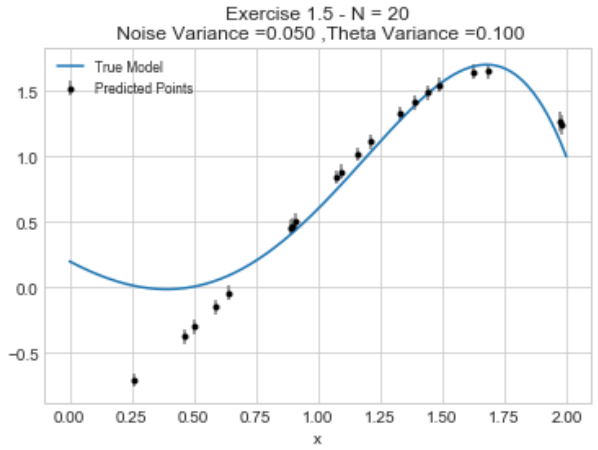


Γράφημα 1.4.4.

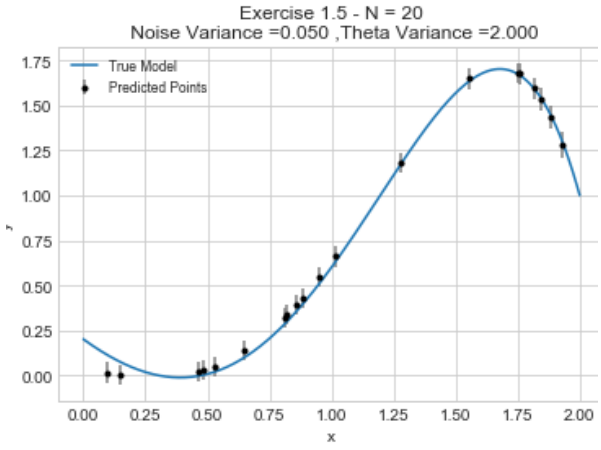
Προβλημα 1.5

Αυτή τη φορά χρησιμοποιηήθηκε και πάλι το σωστό μοντέλο αλλά οι παράμετροι της μέσης τιμής $\boldsymbol{\theta_0}$ της prior κατανομής του **θ** ήταν διαφορετικές από αυτές του πραγματικού **θ** από το οποίο κατασκευάστηκε το training set, συγκεκριμένα $\boldsymbol{\theta_0} = [-10.54, 0.465, 0.087, -0.093, 0, -0.004]^T$. Στα γραφήματα 1.5.1-1.5.4. φαίνονται τα αποτελέσματα της εκτίμησης των respective $\hat{y}$ (μ_y) για τις αντίστοιχες τιμές των 20 τυχαίων x στο διάστημα [0,2], σύμφωνα με τη Bayesian προσέγγιση. Και πάλι με γκρι απεικονίζονται τα error bars για κάθε εκτίμηση $\hat{y}$, τα οποία υπολογίστηκαν βάσει των αντίστοιχων διακυμάνσεων $\sigma_y{}^2$. Δοκιμάστηκαν δυο τιμές για το πλήθος των σημείων του training set, N=20 και N=500. Σε όλες τις περιπτώσεις είχαμε σταθερό $\sigma_\eta{}^2 = 0.05$ αλλά δοκιμάστηκαν δύο διαφορετικά $\sigma_\theta{}^2$ (0.1 και 2.0). Από τα

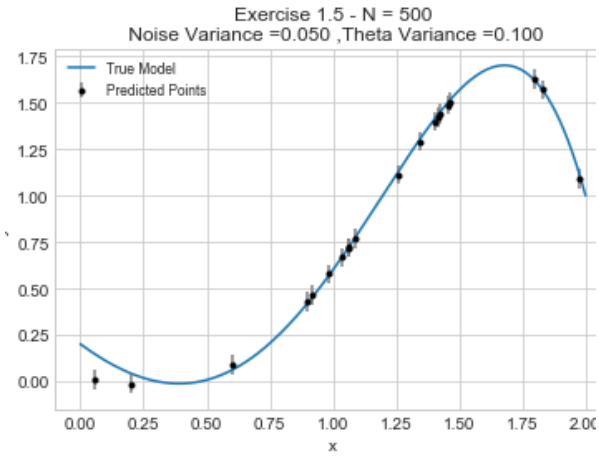
γραφήματα λοιπόν παρατηρούμε ότι καθώς αυξάνεται ο αριθμός των training points οι εκτιμήσεις $\hat{y}$ βελτιώνονται. Όταν επιπλέον, αυξηθεί και η τιμή του $\sigma_\theta{}^2$ οι εκτιμήσεις όχι μόνο βελτιώνονται αλλά προσεγγίζουν πολύ ικανοποιητικά το πραγματικό μοντέλο παρότι ξεκινήσαμε με λάθος εκτίμηση παραμέτρων του $\boldsymbol{\theta_0}$.



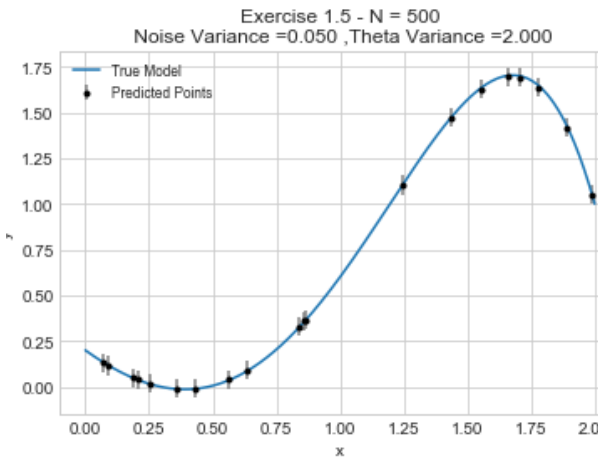
Γράφημα 1.5.1.



Γράφημα 1.5.2.



Γράφημα 1.5.3.

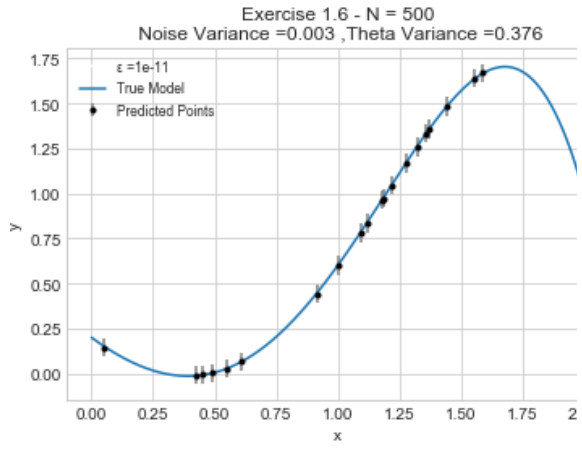


Γράφημα 1.5.4.

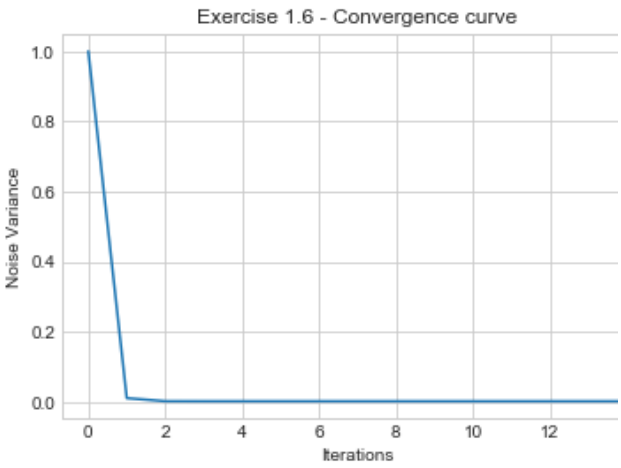
Προβλημα 1.6

Για το πρόβλημα αυτό χρησιμοποιήθηκε ο EM-Algorithm ώστε να εκτιμηθούν οι παράμετροι $\sigma_\eta{}^2$ και $\sigma_\theta{}^2$, οι οποίες θα χρησιμοποιηθούν για την εκτίμηση των παραμέτρων της posterior κατανομής p(θ|y), και στη συνέχεια των $\hat{y}$, ακολουθώντας τη Bayesian προσέγγιση όπως στο πρόβλημα 1.5. Ως αρχικές τιμές δώθηκαν $\sigma_\eta{}^2$=$\sigma_\theta{}^2 = 1$ και η σωστή διάσταση του διανύσματος των παραμέτρων προς προσέγγιση, ενώ δοκιμάστηκαν αρκετές τιμές του ε, ως συνθήκη τερματισμου. Στα γραφήματα 1.6.1 και 1.6.2 παρουσιάζονται τα αποτελέσματα για ε=0.00000000001, ώστε ο αλγόριθμος να μην τερματίσει νωρίς και δώσει αποτελέσματα που αντιστοιχούν σε κάποιο τοπικό μέγιστο. Συγκεκριμένα, στο Γράφημα 1.6.1 φαίνονται α) οι τιμές των $\sigma_\theta{}^2$, $\sigma_\eta{}^2$ που προκύπτουν από τις τιμές τερματισμού του αλγορίθμου a=2.658, b=381.24 β) οι προσεγγίσεις των respective $\hat{y}$ που αντιστοιχούν σε 20 τυχαία $x \in [0,2]$, οι οποίες είναι

πολύ ικανοποιητικές, προσεγγίζοντας πολύ καλά την πραγματική καμπύλη. Στο Γράφημα 1.6.2 φαίνονται οι τιμές του $\sigma_\eta{}^2$ ως συνάρτηση των επαναλήψεων του αλγορίθμου.



Γράφημα 1.6.1.



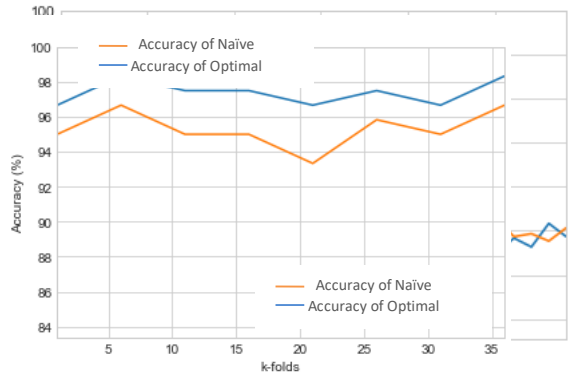
Γράφημα 1.6.2.

Προβλημα 2.2 - 2.3

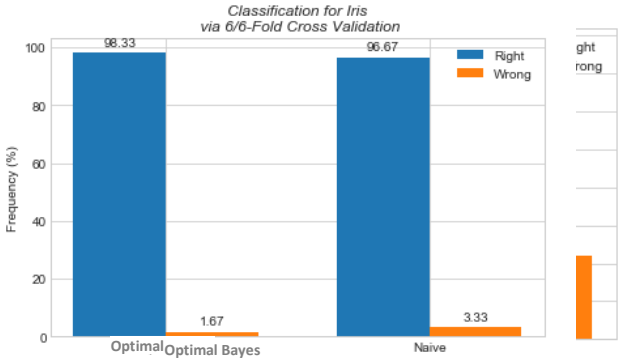
Τα προβλήματα 2.2 και 2.3 δουλεύτηκαν παράλληλα εφόσον το μόνο που αλλάζει ανάμεσα στιον Optimal και τον Naive Bayesian Classifier είναι ο τρόπος υπολογισμού των $p(\boldsymbol{x}|w_i), i = 1, \dots, M$. Συνεπώς απλά με την επιλογή isNaive=True (για περισσότερες πληροφορίες : βλ. σχόλια στον κώδικα) έχουμε $p(\boldsymbol{x}|wi) = p(x_1|wi) \cdot p(x_2|wi) \cdot \dots \cdot p(x_l|wi)$. Αρχικά και στα δύο datasets έγινε optimal και Naive Bayesian Classification με k-fold Cross Validation για διάφορες τιμές του k, ώστε να βρεθεί το k για το οποίο τα ποσοστά των σωστών προβλέψεων γίνονται μέγιστα. Για το data set των pima όπως προκύπτουν απο το Γράφημα 2.2, τα μεγαλύτερα ποσοστά (Γράφημα 2.3) σωστών προβλέψεων του optimal Bayesian Classifier δίνονται για 171-Fold Cross Validation ενώ του Naive για 91-Fold Cross Validation. Αντίστοιχα, για το data set των iris όπως προκύπτουν απο το Γράφημα 2.4, τα μεγαλύτερα ποσοστά (Γράφημα 2.4) σωστών προβλέψεων του Naive Bayesian Classifier δίνονται για 6-Fold Cross Validation ενώ του Naive για 6-Fold Cross Validation. Επιπλέον, υπολογίστηκε η μέση τιμή των σωστών απαντήσεων για τα διάφορα k και ως μέτρα precision οι τυπικές αποκλίσεις και οι απόλυτες τυπικές αποκλίσεις όπως φαίνονται στον Πίνακα 1. Από τα αποτελέσματα αυτά και από τα γραφήματα παρατηρούμε ότι στην περίπτωση των Iris data αποδίδει καλύτερα ο Optimal ενώ στην περίπτωση των Pima τα ποσοστά είναι καλύτερα για τον Naive. Στην περίπτωση των Iris, ο Bayes Classifier ειναι και accurate και precise, με τον Optimal να αποδίδει σταθερά καλύτερα από τον Naive, πιθανότατα λόγω της σχετικά μικρής διάστασης του πίνακα των χαρακρηριστικών. Αντίθετα, στην περίπτωση των Pima τα ποσοστά των σωστών ταξινομήσεων δεν είναι σταθερά, εφόσον και για τον Optimal και για τον Naive οι δείκτες precision είναι σχεδόν διπλάσιου απο τους αντίστοιχους δείκτες για το Iris dataset. Αυτό σημαίνει πως ο Bayes Classifier ίσως δεν είναι κατάλληλος ο κατάλληλος ταξινομητής για το dataset των Pima. Αυτό ίσως έχει να κάνει με τον αριθμό των χαρακτηριστικών που έχει το data set των Pima σε σχέση με τον αριθμό των data points, δηλαδή το curse of dimensionality.

	Optimal			Naive		
	Mean	**STD**	**Absolute STD**	**Mean**	**STD**	**Absolute STD**
Pima	73.72	1.3	1.1	74.93	1.14	1.31
Iris	97.39	0.65	0.54	95.32	0.65	0.54

Πίνακας 1

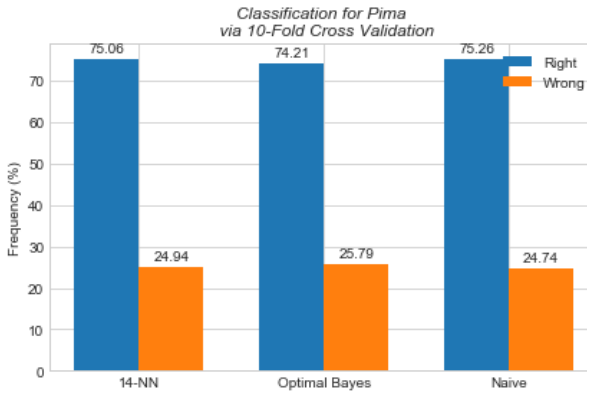


Γράφημα 2.4.

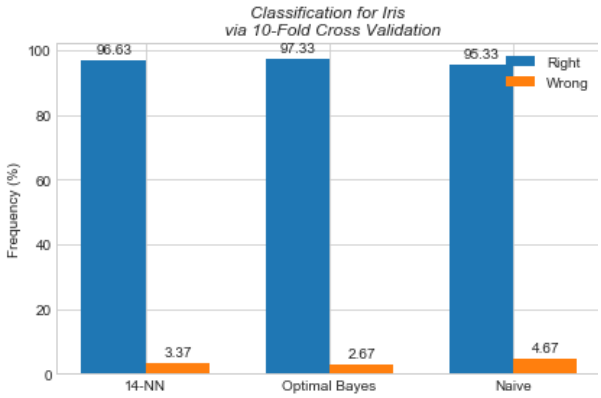


Γράφημα 2.5.

Για τη σύγκριση των αποτελεσμάτων μεταξύ των τριών τώρα classifiers των ερωτημάτων 2.1, 2.2, 2.3 χρησιμοποιήθηκε Ten-Fold Cross Validation, και k=14 στον k-NN. Τα αποτελέσματα φαίνονται στα Γραφήματα 2.5, 2.6.



Γράφημα 2. 6.



Γράφημα 2. 7.

<u>Οδηγός για τον κώδικα</u>

Προβλημα 1.1 : 1_1_Linear_Regression.py

Προβλημα 1.2 : 1_2_Linear_Regression.py

Προβλημα 1.3 1_3_Ridge_Regression.py

Προβλημα 1.4, 1.5 : Bayesian_Inference_1_4_5.py

Προβλημα 1.6 : EM_Algorithm

Προβλημα 2.1 :

Προβλημα 2.2,2.3 : Bayesian_Classifier_2_2_3.py

compare.py

Προβλημα 2.4 :

<u>Βοηθητικά</u> :

Polynomial.py

Module.py