**תרגיל 2 random forest and random regression forest**

**תאריך הגשה 7/1/2016**

בקורס למדנו על עצי החלטה ועצי רגרסיה. בתרגיל זה תתבקשו לממש את random forest לשני המקרים האלה לסיווג ולהערכת משתנה רציף (רגרסיה).

לגבי סיווג אנא השתמשו בפונקציית ג'יני לצורך בחירת הפיצול האופטימלי ובמזעור שונות לצורך בחירת פיצול לצורך רגרסיה.

התוכנית צריכה להיות כללית: מספר השדות ומספר המשתנים יכול להשתנות. מספר המחלקות אף הוא יכול להשתנות וערכיו נעים בין 0 למספר המחלקות פחות 1.

בקובץ הקלט שאותה התוכנית שלכם תקרא יופיע בשורה הראשונה מספר השדות ומספר המחלקות. במקרה של רגרסיה מספר המחלקות יהיה 0.

הדוח שלכם יתחיל עם תיאור בין כשני עמודים של התוכנה שבניתם (מחלקות עיקריות פונקציות מעניינות ומה מעניין לדעתכם בפתרון שלכם).

תקבלו שני קבצי קלט ל random forest. כל קלט יינתן פעמיים. פעם כקלט לתוכנית שלכם ופעם כקלט ל weka. תקבלו קובץ train וקובץ test. הריצו את weka על שתי הקלטים ודווחו על התוצאות שקבלתם והשוו לתוצאות שהתוכנית שלכם קיבלה על אותם קלטים.

הקלט הראשון יכלול נתונים דו-ממדיים כך שתוכלו (ותצטרכו) להגיש גרף שבו רואים את הנקודות כשכל נקודה מסומנת בצבע של המחלקה שלה והאם האלגוריתם הצליח לסווג אותה נכון. ייצרו את הגרף גם ל train וגם ל test.

לגבי הקלט השני ייצרו גרף שבציר ה x שלו נתון מספר העצים בין 1 ל 200 ביער שבניתם ובציר ה y שלו השגיאה של היער לגבי קובץ ה train וה test.

ייצרו גרפים כאלה גם עבור שני המקרים של הרגרסיה.

לגבי הדוגמא הגדולה של הסיווג והדוגמא הגדולה של הרגרסיה דרגו את השדות של הקלט לפי החשיבות שלהם בבניית המסווג. הסבירו איך דרגתם את השדות.

הריצו את כל ארבעת הדוגמאות גם ב weka והסבירו את ההבדלים אם קיימים בין הביצועים של האלגוריתם שכתבתם לביצועים של weka. יש לצפות שלא יהיו הבדלים גדולים.

בונוס: לגבי הגרפים שייצרתם לדוגמת הסיווג הגדולה ודוגמת הרגרסיה הגדולה ייצרו את אותם הגרפים עבור התוצאה ש weka ייצרה עבור הבעיה הזאת. לצורך זה תצטרכו לגשת למבני הנתונים ש weka מייצר תוך שימוש ב api שלו.

קבצים:

1. data.txt data2.txt data.arff data2.arff: נתונים ממימד 2 של שלוש מחלקות.
2. forestS.txt forestS2.txt forestS.arff forestS2.arff נתונים ממימד 49 של 4 מחלקות.
3. sphere.txt sphere2.txt sphere.arff sphere2.arff נתונים ממימד 10 ומשתנה שעליכם לחזות. במקרה זה הערך שעליכם לחזות הוא סכום ריבועי הערכים הנתונים. אין להשתמש בזה בתרגיל זה רק לידיעה.
4. forestR.txt forestR2.txt forestR.arff forestR2.arff נתונים ממימד 14 ומשתנה מסוג real שעליכם לחזות.

**ניקוד וציון:**

ניתן לקבל יותר מ 100 נקודות עבור התרגיל.

התרגיל ינוקד על פי הקריטריונים הבאים:

1. נכונות.
2. בהירות
3. יעילות המימוש
4. כלליות (ניתן להשתמש במבנה הנתונים עבור נתונים מסוגים שונים).
5. אלגנטיות המימוש
6. ביצוע ניתוח המשתנים המעניינים על הפתרון שנוצר על ידי weka.

בהצלחה,

צוות הקורס.