⚡️🚀המאמר היומי של מייק 31.08.24: ⚡️🚀

Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review

היום סוקרים מאמר לא רגיל - קודם כל זה מאמר סקירה בעצמו והוא לא מאוד טרי (מלפני כמעט שנתיים). המאמר בנושא של explainability של מודלי למידת מכונה. רוב מודלי ML היום הם רשתות נוירונים מאוד עמוקות ולרוב הם נשארים בתור קופסא שחורה עבורנו - מחקרי explainability מנסים לשפוך אור על ״מה שקורה בתוך הקופסא השחורה הזו״.

המאמר הזה נותן סקירה של אחד הפרדיגמות העיקריות המשמשות למחקר explainability של מודל ML - ניתוח counterfactual. כלומר חוקרים מה צריך לשנות בדגימה(איזה פיצ'רים) כדי שהיא תסווג לקטגוריה (קלאס) אחרת ועל ידי כך נבין יותר טוב למה המודל סיווג את הדוגמא המקורית לקטגוריה המקורית. דרך אגב יש שיטות explainability שחוקרות את המודל בצורה אחרת. למשל קיימות שיטות שמנסות לקרב את המודל המורכב על ידי מודל פשוט יותר (עץ או רגרסיה לינארית) בטווח מסוים של דוגמאות. שיטות נוספת מנתחות המנסות להסביר את חיזויו של המודל לדוגמא ספציפית.

אז מה בעצם חשוב לנו מאוד בשיטות counterfactual? קודם כל חשוב לנו לשנות כמה שפחות פיצ'רים של הדוגמא הנחוצים ל״העברתה״ לקטגוריה אחרת וגם השינוי בפיצ'רים אלו צריך להיות די קטן כדי להבין את ״מבנה גבול״ בין הקטגוריות השונות מבחינת המודל. השני ולא פחות חשוב השינוי הזה צריך להיות ״חוקי״ כלומר הדוגמא הנוצרת צריכה להיות הגיונית וולידית (כלומר שטח הבית לא יכול להיות שלילי). בנוסף השינוי בדוגמא צריך לעבור במסלול הגיוני כלומר בקרבה של הדוגמאות האחרות מהדאטהסט. וכמובן יש עוד דרישות לשינוי שאנו מחוללים לדוגמא כדי להפוכה ל-counterfactual.

וכל הפרטים המעניינים במאמר כמובן…

https://arxiv.org/abs/2010.10596