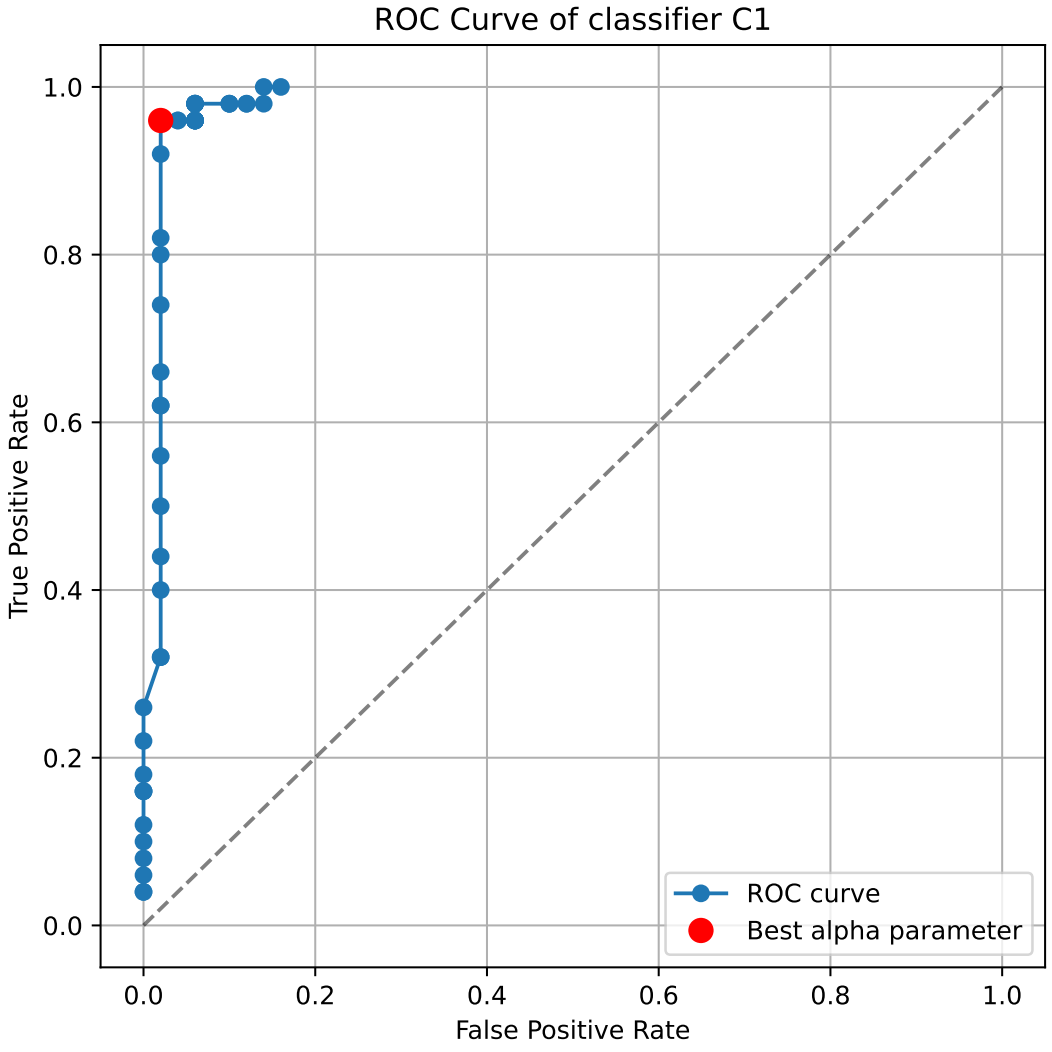
Výběr klasifikátoru

1. Výběr vhodného parametru

Než začneme vůbec hledat nejlepší parametr pro klasifikátor C_1, podíváme se, jaké situace obsahuje soubor s ground-truth. Vídíme, že počet pozitivních a negativních situací je dostatečně velký (50) a zároveň se počty shodují. Máme tedy relevantní situace k tomu, aby mělo smysl hledat samotný parametr.

Abychom mohli dále pokračovat, pro jednotlivé parametry α spočítáme množství TP, TN, FN, TN případů ze srovnání ground-truth a výsledků klasifikátoru C_1. Z těchto tříd vypočítáme recall (TPR) a FPR, které vykreslíme do obrázku 1. Když spojíme takto vzniklé body ve všech dostupných hodnotách α, získáváme ROC křivku.

Bod obecně nejlepšího α hledáme jako bod nejblíže (ve smyslu eukleidovské vzdálenosti) k bodu $(0, 1)$, kde by byl klasifikátor optimální, tedy že by určil všechny pozitivní případy správně a zároveň by nikdy neklasifikoval jako pozitivní to, co pozitivní není. Zde jsme našli 3 nejbližší body na stejných souřadnicích, a to $\alpha = \alpha_{22}$, $\alpha = \alpha_{23}$, $\alpha = \alpha_{24}$.



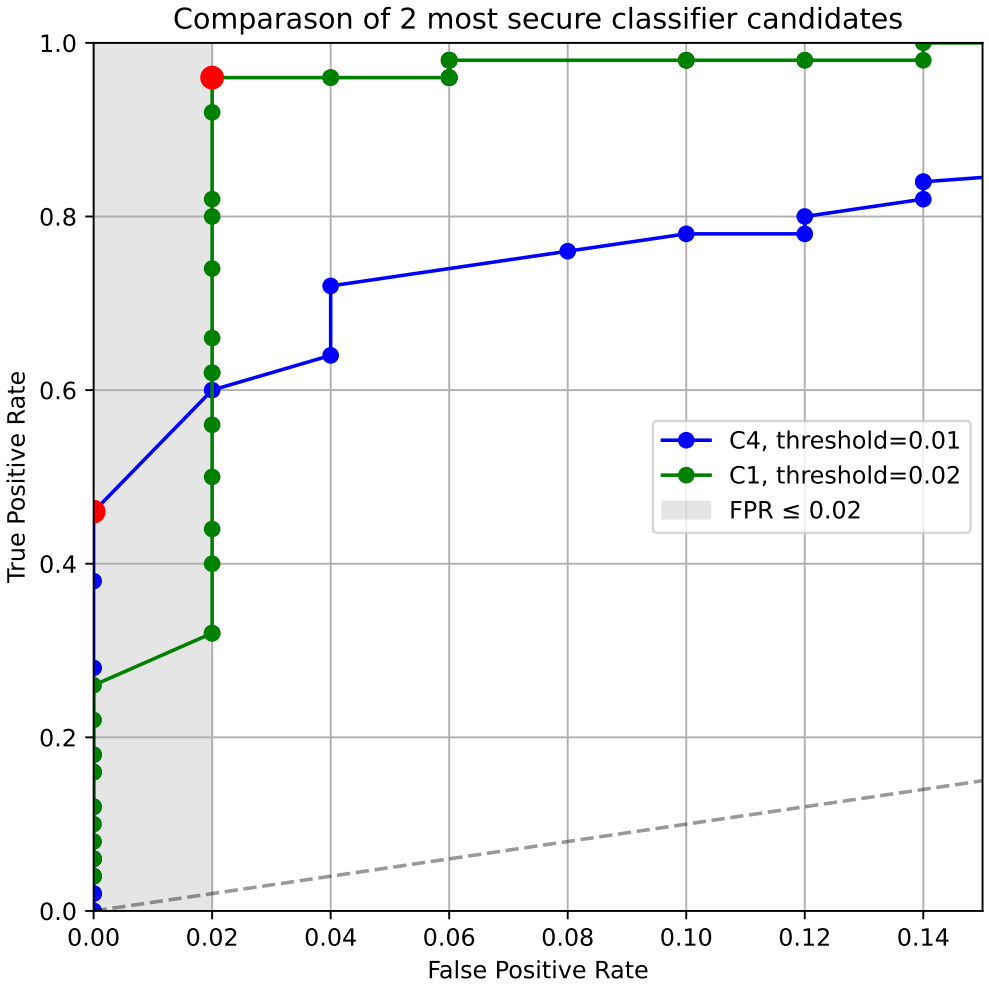
Obrázek 1: ROC křivka klasifikátoru C_1

2. Přísně tajné

V této části máme díky dostatku času možnost opakovat načtení otisku prstu vícekrát. To nám mírně mění nároky na náš klasifikátor, protože nám to umožňuje minimalizovat FPR částečně i na úkor TPR, což omezí riziko neoprávněného přístupu k tajným dokumentům. Použijeme metriku, která bude uvnitř pravého ϵ-okolí nulového FPR hledat maximum rozdílu $TPR - FPR$.

Při určování jsem vzhledem k citlivosti dat zvolil $\epsilon := 0.01$, pro které mi vyšel nejlepší klasifikátor C_4 s $\alpha = \alpha_{11}$, kde $TPR = 0.46$, což sice znamená nutnost přikládat prst ke čtečce několikrát, ale riziko průniku cizího uživatele je minimální. Na druhou stranu bude ale klasifikátor náchylný k malým změnám struktury otisku, takže např. agent s odřeným prstem bude mít výrazně nižší šanci dostat se do vlastního trezoru.

Posuneme-li hranici na $\epsilon := 0.02$, což je stále přijatelná hodnota, bude metrice vyhovovat lépe klasifikátor C_1 s $\alpha = \alpha_{22}$ a $TPR = 0.94$, tentokrát s menším rizikem, že se agent nedostane do svého vlastního trezoru. Ale vzhledem k tomu, že je lepší variantou zabezpečená data zničit, než aby se k nim, byť s malou šancí dostal někdo jiný, raději zvolíme první variantu.



Obrázek 2: Porovnání ROC křivek klasifikátorů C_1 a C_4

3. Hlavně bezpečně

Pro určení, jestli je cizí klasifikátor lepší než náš nejlepší, použijeme metriku popsanou v předchozím úkolu. V přiloženém souboru se jedná o funkci *is_better()*.