Zadání Cvičení #9

Popis dat: Pracovní data jsou uložena v souboru data.csv, který je k dispozici ke stažení na Moodle stránce tohoto předmětu, ve složce příslušného cvičení. Pro načtení dat do Matlabu využijte funkce `readtable`.

Data jsou ve formátu tabulky, která obsahuje data od pacientů s Parkinsonovou nemocí (label **PD**) a data od kontrolní skupiny zdravých lidí (label **HC**). Tabulka obsahuje ID kódy subjektů, identifikátory příslušnosti ke skupině (labely), údaje o pohlaví, věku, a hodnoty těchto dvou akustických parametrů:

- **stdPWR** – Směrodatná odchylka intenzity řeči, jednotka dB.
- **stdF0** – Směrodatná odchylka F0, jednotka půltóny – *semitones* (st).

Oba parametry byly vyhodnoceny z nahrávek souvislé řeči (monologu) pomocí programu PRAAT (viz Boersma and Weenink 2017).

Zadání úlohy	body
Vizualizujte si data pomocí 2D-scatter grafu. Obě skupiny od sebe **barevně odlište** a použijte libovolné značky a velikost tak, aby byly data **dobře viditelná**. Navíc si můžete případně vykreslit rozdělení dat pomocí boxplotů. **Odpovězte svými slovy na tyto otázky:** a) Pohledem na vaší vizualizaci zhodnoťte, jak dobře lze oddělit skupiny v 2D prostoru pomocí lineární rozhodovací hranice. Do vizualizace můžete zakreslit (od ruky) kam byste rozhodovací hranici umístili. b) Zhodnoťte, zda není pro klasifikaci některý z parametrů nadbytečný, a zvažte, jak by se obecně daly nadbytečné (redundantní) parametry automaticky identifikovat ve vícedimenzionálních datasetech.	1
Proveďte klasifikaci dat pomocí následujících metod a pro každou z nich vykreslete do obrázku s daty příslušnou rozhodovací hranici (*decision boundary*), abyste je mohli mezi sebou snadno vizuálně porovnat: 1. **Logistická regrese** (*Logistic regression*) • Použijte funkce `glmfit` a `glmval`, se standardním nastavením pro binomiální rozdělení výstupní funkce. • Rozhodovací hranici můžete vypočítat z definice logistické funkce nebo můžete využít funkci `contour` na úrovni `0.5`. 2. **Lineární diskriminantní analýza** (LDA) • Pro výpočet LDA modelu použijte funkci `fitdiscr`. • Klasifikaci dat můžete provést pomocí funkce `predict`. • Rozhodovací hranici vykreslete za použití koeficientů z vypočteného LDA modelu.	1.5

3. **Support Vector Machine** (SVM) s lineárním jádrem • Výpočet SVM modelu proveďte pomocí funkce `fitcsvm`. • Pokud chcete, můžete si vykreslit vhodným způsobem do obrázku jaké datové body použil model jako *Support Vectors*. • Koeficienty pro vykreslení rozhodovací hranice SVM modelu získáte z Matlabovské proměnné (*structure*) modelu (pole `Beta` a `Bias`).	
Porovnejte jednotlivé klasifikátory z hlediska výpočetní náročnosti a přesnosti klasifikace. a) Pro měření výpočetní náročnosti využijte časoměrné funkce `tic` a `toc`. b) Pro každý z klasifikátorů vypočtěte a zobrazte matici záměn (*Confusion matrix*) – využijte funkcí `confusionmat`, resp. `confusionchart`. Přesnost klasifikace určete jako procento dat, které by model přiřadil do správné skupiny. • **Zapište časy výpočtu pro všechny modely a procentuální přesnosti klasifikace.** • **Svými slovy zhodnoťte výsledky a porovnejte modely.**	1.5
## Nepovinný bonus: Implementujte manuálně algoritmus pro výpočet lineární diskriminační analýzy **LDA**. Pro otestování funkce použijte data uložená v souboru bonus.csv, který je dostupný ke stažení na Moodlu (proměnné jsou značeny *x* a *y*). **Odevzdejte obrázek s vaší kódovou implementací a grafické zobrazení dat a rozhodovací hranice.**	0.5

Reference

Boersma, P. and Weenink, D. (2017). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.30, retrieved 22 July 2017 from http://www.praat.org/