Prov i Maskininlärning VT24

Ansvarig lärare: Raphael Korsoski

Max 23p. Godkänt 10p. Väl Godkänt 17p.

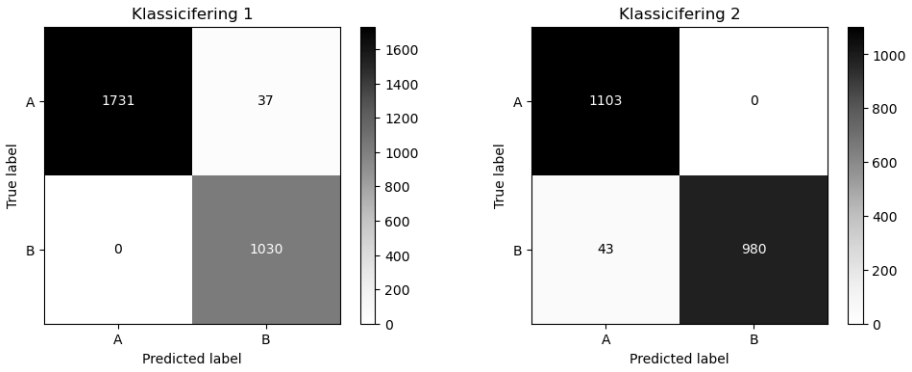
Tillåtna hjälpmedel: Miniräknare.

Datum: 5e april, 2024

Tid: 9.00 - 13.00

Alla svar skall motiveras för full poäng. *Lycka till!*

1 Betrakta följande diagram.



a) Vilken av "Klassificering 1" eller "Klassificering 2" har högst **precision** med avseende på klass **B**? *Tips: precision* = $\frac{TP}{TP+FP}$

1p

b) Vilken av "Klassificering 1" eller "Klassificering 2" har högst **recall** med avseende på klass **B**? *Tips: recall* = $\frac{TP}{TP+FN}$

1p

c) Vid en raketuppskjutning som delvis flyger högt över städer används ett system som tolkar **B** som "aktivera självförstörelsemekanismen", vilket undviker att raketen dimper ner i bebyggelse om något går fel. Vilken klassificerare är mest lämplig i detta fall? Motivera ditt svar.

2p

Var god vänd

2 Givet följande linjära modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$
$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 1 & 3 & 2 \end{bmatrix}^T$$

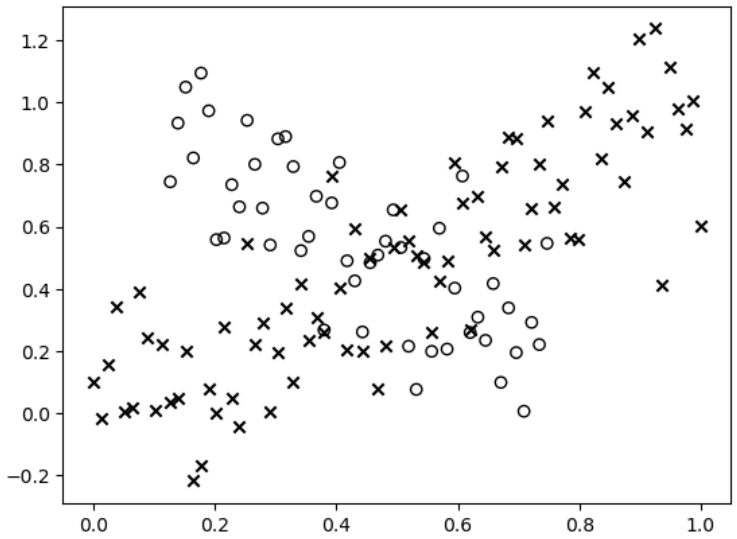
a) Förutsäg värdet för punkten $(1, 2)$

1p

b) Låt $f(x_1, x_2) = \mu_{Y|\mathbf{X}}$. Är f linjär? Motivera ditt svar.

2p

3 Betrakta nedanstående figur



a) Är datamängden linjärt separerbar? *Tips: Tänk på antalet dimensioner*

2p

b) Längs vilka riktningar finns de två första principalkomponenterna?

2p

Var god vänd

4 LASSO och Ridge regression använder l_1 respektive l_2 norm. Förklara tendenserna för dessa vad gäller koefficienterna i ekvationssystemet $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$.

2p

5 En utvecklare använder en kostnadsfunktion

$$\boldsymbol{C}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y})^2 + \sum_{j=1}^{p} \beta_j^2$$

Utvecklaren tycker att regressionen får oväntat dåligt resultat jämfört med förväntningarna från tidigare samlad statistik och dessutom lägger för stor vikt vid vissa parametrar.

Föreslå två förbättringar.

3p

6 Ge två exempel på när en *train/test* split **inte** är relevant.

2p

7 Ett företag har valt att använda ett beslutsträd i ett felrapporteringssystem. Beslutsträdet avgör vilken avdelning felrapporten skall skickas till. Trots att det bara finns fem avdelningar blir det ett träd med djup 10 och flera hundra löv. Dessutom måste databasuppslagninar göras för att testa villkoren i grenarna. Systemet blir väldigt långsamt och kan inte utnyttja parallellism.

Föreslå en bättre algoritm på samma data som kräver så liten ändring som möjligt.

2p

8 Under utvecklingen av en klassificeringsalgoritm vill en utvecklare göra en *polynomexpansion* på designmatrisen medans en annan vill byta till en *polynomkernel* istället. Vilket sätt är att föredra? Motivera ditt svar.

3p