Gedankenprotokoll Statistisches Lernen Klausur

09.02.23

1. Aufgabe: Grundlagen der Wahrscheinlichkeitsrechnung

- **Kartenspiel**: 52 Karten (2-10, Bube, Dame, König, As von 4 Farben)
- **Gewinne**:
    - As: 100$
    - Karten > 10 (außer As): 10$
    - Alle anderen Karten: -6$
- **Berechnungen**:

1. Erwartungswert des Gewinns pro Spiel
2. Varianz und daraus Standardabweichung berechnen

---

2. Aufgabe: Grundlagen Statistisches Lernen

- **Diagramm**:
    - **x-Achse**: Flexibilität des Modells (niedrig → hoch)
    - **y-Achse**: 5 Kurven:
        - Bias
        - Variance
        - Training Error
        - Test Error
        - Non-reducible Error
- **Begründung**:
    - Bias: Sinkt mit steigender Modellkomplexität
    - Variance: Steigt mit steigender Modellkomplexität

- Training Error: Sinkt mit steigender Modellkomplexität
- Test Error: U-förmig, Minimum bei moderater Flexibilität
- Non-reducible Error: Bleibt konstant

---

3. Aufgabe: Lineare Regression

- **Gegebene Regressionsgeraden**:
  - fa(x)=40−0,36·xf_a(x) = 40 - 0,36 \cdot xfa(x)=40−0,36·x
  - fb(x)=45−0,4·xf_b(x) = 45 - 0,4 \cdot xfb(x)=45−0,4·x
- **Berechnungen**:

1. Restfehler jedes Punktes (RSS)
2. Bestimmen, welches Modell (f_a oder f_b) besser passt basierend auf RSS

- **Standardabweichung von beta_1**:
  - Bedeutung des Intervalls (Mittelwert ± 2 Standardabweichungen)
  - Wie kann die Standardabweichung von beta_1 durch Regularisierung reduziert werden?

---

4. Aufgabe: Regularization

- **Lasso vs. Ridge Regression**:
  - **Lasso**: L1-Regularisierung, kann Koeffizienten auf Null setzen
  - **Ridge**: L2-Regularisierung, schrumpft Koeffizienten, setzt sie aber nicht auf Null

---

5. Aufgabe: Nicht-Lineare Regression

- **Beispiele für Generalized Linear Models (GLMs)**:
  1. Poisson-Regressionsmodell
  2. Logistische Regression
  3. Negative Binomial Regression
- **Mathematische Repräsentation (logistische Regression)**: P(Y=1|X)=11+e−(β0+β1X)P(Y=1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}P(Y=1|X)=1+e−(β0+β1X)1

---

6. Aufgabe: Entscheidungsbäume

- **Optimierungsparameter**:
    - Gini-Index
    - Informationsgewinn
- **Aufgaben**:

1. Entwurf des Entscheidungsbaums für das gegebene 2D-Diagramm
2. Zuordnung eines gegebenen Punktes (X1, X2) und Bestimmung des Y-Werts

---

7. Aufgabe: k-fold Cross-Validation

- **Training/Test-Samples**:
    - Daten werden in k Teile unterteilt
    - Jeder Teil wird einmal als Testset und k-1 mal als Trainingsset verwendet

---

8. Aufgabe: Unsupervised Learning (PCA)

- **Szenario**:
    - Zehnkampf mit 10 Disziplinen
    - PCA-Diagramm mit 2 Hauptkomponenten
- **Berechnungen**:

1. Prozentuale Varianz, die nicht durch die 2 Hauptkomponenten erklärt wird
2. 2 Sportarten, deren Ergebnisse stark korrelieren
3. Sportarten, die mit PC1 und PC2 stark korrelieren
4. Quadrant, in dem Werf-Sportarten fallen (I-IV)

---

9. Aufgabe: Multiple Hypothesis Test

- **Gegebener R-Code**:
    1. Erstellen einer 10 x 50 Matrix mit normalverteilten Werten (Mittelwert 0, Standardabweichung 1)
    2. Verschieben von 25 Spalten auf Mittelwert 2
    3. Durchführung eines Hypothesentests, um festzustellen, ob der Mittelwert = 0 ist

- **Ergebnisse**:
    - Bestimmung der Multi-Hypothesen-Test-Methode (z.B. Benjamini-Hochberg)
    - Bestimmung des relevanten Parameters (FWER oder FDR)
    - Anzahl der abgelehnten Hypothesen