**Methods & Alogrithm Zusammenfassung Theorie**

1. **Lineare Regression**

* eine der am meisten genutzten statistischen Methoden
* drückt die mathematische Beziehung zwischen zwei Variablen aus
* Annahme: lineare Zusammenhang zwischen Antwort- und Prädiktorvariable
* Linearität vereinfacht möglichen tatsächlichen
* gesucht: Modell für Zusammenhang zwischen Prädiktor- (x) und Antwortvariable (y), d.h. finde beste Werte B0, B1 für gegebene Punkte (xi, yi)

warum lineare Regression…?

* um Beziehungen besser zu verstehen / besser beschreiben zu können
* etwa: gibt es einen Zusammenhang zwischen Anzahl der Freunde in sozialen Medien und der Zeit, die eine Person täglich auf derartigen Plattformen verbringt

Ein Bild, das Text, Schrift, weiß, Handschrift enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Schrift, weiß, Design enthält.

Automatisch generierte BeschreibungGerade f(x) liegt optimal zu allen Punkten (xi, yi) <-> vertikale Abstände (genauer: Quadrat der Abstände) zwischen den Punkten und der Regressionsgeraden minimal

Berechnung des Fehlers: mittlere quadratische Abweichung

**MSE** ist ein gutes Mass für die Varianz (vgl. Varianz aus der Statistik), d.h. wie stark die

vorhergesagten Werte (yi) von den tatsächlichen / beobachteten Werten (yi) variieren

Vorsicht: je kleiner der **MSE**, desto kleiner der Fehler, aber je kleiner **R2**, desto höher die Variabilität (🡪 desto schlechter die Güte der Regressionsgeraden)

Bestimmtheitsmass **R2** drückt Anteil der Variabilität in den Messwerten aus

**#test\_size** Anteil Testfälle %,

**random\_state** bestimmt zufällige Durchmischung der Daten vor dem Split, Daten jedes Mal, wenn der Code ausgeführt wird, auf die gleiche Weise aufgeteilt werden. Das ist nützlich, um die Ergebnisse reproduzierbar zu machen, da dieselbe Aufteilung der Daten bei jedem Ausführen des Codes beibehalten wird.

1. **K-Nächste / k-Mitten**

* Punktewolke von bereits klassifizierten Objekten (z.B. Krebsrisiko bei Patienten)
* neue, unklassifizierte Patienten sollen aufgenommen und klassifiziert werden **(kategoriale Variablen)**

**Ein Bild, das Reihe, Text, Diagramm, Schrift enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Schrift, weiß, Typografie enthält.

Automatisch generierte BeschreibungEin Bild, das Reihe, Text, Diagramm, Schrift enthält.

Automatisch generierte BeschreibungAbstand:**

Euklidischer Abstand

Jaccard'scher Abstand / Ähnlichkeit (Distanz zwischen einer Menge von Objekten

Ein Bild, das Text, Schrift, Screenshot, weiß enthält.

Automatisch generierte BeschreibungHamming Abstand Anzahl der Stellen, an denen sich zwei gleich lange Sequenzen unterscheiden

Cosinus Ähnlichkeit (beschreibt Ähnlichkeit zweier Vektoren)

Manhattan Abstand (Abstand zweier k-dimensionaler Vektoren)

**Klassifizierung (Mehrheitsentscheid:** Berechnung Distanzen, k-Nachbarn finden, Mehrheit der Kategorie entscheidet

**Bewertungsmass:**

* Treffsicherheit, berechnet Anteil resp. Anzahl korrekter Vorhersagen
* Hamming Verlust (Fehlinterpretationsrate), berechnet durchschnittliche(n) Distanz / Verlust zwischen zwei Mengen

**Nächste Nachbarn (kNN):** Wir wollen Objekte einer Klasse zuordnen und orientieren uns an Objekten in der Nähe, die bereits einer Klasse zugeordnet sind. Wir müssen nur k, also «den k nächsten Nachbarn» bestimmen und deren Klasse übernehmen. Betrachen ähnliche Objekte, Mehrheitsentscheid 🡪 bereits Klassifizierte Daten (Training), supvervised Learning

**K-Means:** Zwischen einzelnen Datenpunkten eines Clusters gibt es eine Mitte (Zentroid/Means). Bei hinzufügen von neuen Datenpunkten "wandert" der Zentroid. K-Means funktioniert nur gut, wenn die Cluster räumlich gut getrennt sind und ähnlich gestreut sind und die Streuung kugelförmig ist. Im Grenzbereich der Cluster kommt es zu Falschzuweisungen aufgrund der Entfernungseinschätzung (Konvergenz nicht garantiert) 🡪 Unsupvervised Learning, mit unklassifizierten Daten, Muster selber erkennen

Ein Bild, das Text, Reihe, Diagramm, Steigung enthält.

Automatisch generierte Beschreibung

Das Ellbow-Kriterium zur Bestimmung der passenden Clusterzahl: Man probiert mehrere Cluster-Anzahlen aus und schaut wie sich die Summe der Entfernungen zu den Zentroiden entwickelt. Dort wo die Kurve nur noch linear sinkt, ist der Ellenbogen und die passende Zahl.

1. **Hauptkomponentenzerlegung (PCA)**

Zu viele Attribute je Datenpunkt 🡪 Fluch der Dimension

viele Merkmale annähernd konstant / in hohem Masse miteinander korreliert, d.h. Daten liegen innerhalb eines niedrig dimensionalen Subraums

**Principal Component Analysis (PCA)** findet diejenige Hyperebene, die den Daten am nächsten ist und projiziert sie darauf, d.h. mittlerer quadratischer Abstand zwischen Daten und Hyperebene ist minimal. grösste Varianz bleibt erhalten (🡪Informationsverlust)

2D 🡪 Gerade 3D 🡪 Ebene nD 🡪 (n-1)-dimensionale Ebene, also eine Hyperebene

Bild besteht aus niedrig-frequentem und hochfrequentem Anteil.

Das "Rauschen" oder "Noise" bei Bildern bezieht sich auf unerwünschte oder zufällige Variationen in den Pixelwerten eines Bildes. Es tritt häufig als Folge von Störungen während der Bildaufnahme, Übertragung oder Speicherung auf. Rauschen kann die Bildqualität beeinträchtigen und die visuelle Wahrnehmung oder die Genauigkeit von Bildverarbeitungsalgorithmen negativ beeinflussen. Diese Methoden zielen darauf ab, das Rauschen zu reduzieren, während wichtige Bilddetails und Kanten erhalten bleiben.

1. **Lineare Diskriminanzanalyse (LDA)**

Dimensionsreduktion

robuste Klassifizierung (d.h. selbe Ergebnisse mit oder ohne Normalisierung der Daten)

* Methode des überwachten Lernens
* Ausgangspunkt: klassifizierte Daten mit d Merkmalen in k Klassen
* Diskriminanzfunktion teilt (Sub)Raum in k zugehörige disjunkte Regionen
* einfache Klassifikation: prüfe, in welcher Region das Datum liegt
* Diskriminanzanalyse setzt Normalverteilung der Daten voraus
* simple Annahme: n Beobachtungen (mit d Merkmalen), die sich alle mindestens einer von k Klassen zuordnen lassen
* Ein Bild, das Reihe, Diagramm, Steigung enthält.

  Automatisch generierte Beschreibunggeometrisches Verfahren: Beobachtungsraum wird durch eine Reihe von Hyperebenen derart partitioniert, dass diese die Grenzen zwischen jeweils zwei Klassen darstellen

1. Projektion von Punkgen auf eine Gerade
2. Punkt hat Abstand auf der Geraden durch V von Ursprung

**Linear Discriminant Analysis (LDA)**

Ein Bild, das Diagramm, Reihe, Design enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Diagramm, Kreis, Darstellung enthält.

Automatisch generierte BeschreibungProjektion d-dimensionaler Daten in einen Unterraum (<d), sodass Variabilität zwischen den Klassen maximal und Variabilität innerhalb einer Klasse minimal wird

finde optimales w, sodass Vk1, Vk2 minimal und Vk1k2 maximal

mit optimaler Projektionsebene (aufgespannt durch w) ist auch Diskriminanzfunktion

bekannt 🡪 Klassifikation durch Zuordnung: Punkt x gehört zu Klasse Ki falls x in Region Ri

|  |  |
| --- | --- |
| **LDA** | **PCA** |
| Varianz / Streuung minimieren | Varianz / Streuung erhalten |
| Aus Originaldaten nur Klassen in dimensionalem Raum möglichst getrennt | Originaldaten möglichst behalten in einem reduzierten Raum |

1. **Tensor**

Tensoren sind mehrdimensionale Reihungen (Arrays) und damit eine Generalisierung von

Matrizen in höhere Dimensionen

Ein Bild, das Reihe, Entwurf, Design, Muster enthält.

Automatisch generierte BeschreibungTensorfasern

* Entstehung durch 'Festhalten' aller Indizes bis auf einen
* höherdimensionales Analogon zu Zeilen / Spalten einer Matrix

Ein Bild, das Design, Reihe enthält.

Automatisch generierte BeschreibungTensorschichten

* zweidimensionaler Ausschnitt eines Tensors
* Entstehung durch 'Festhalten' aller Indizes bis auf zwei

Matrifizierung (Entfaltung / Ebnung eines Tensors)

* Idee: Umordnung der Elemente eines N-Mode Tensors in eine Matrix

Ein Bild, das Text, Schrift, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Diagramm, Schrift, Screenshot enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Schrift, Screenshot, Zahl enthält.

Automatisch generierte Beschreibung

Orthogonal, wenn Skalarprodukt =0

Ein Bild, das Schrift, Text, Handschrift, weiß enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Schrift, Diagramm, Quittung enthält.

Automatisch generierte Beschreibung

Skalarprodukt

Matrize (m Zeilen, n Spalten)

Ein Bild, das Text, Schrift, Typografie enthält.

Automatisch generierte BeschreibungTransponiert (n Spalten, m Zeilen)

Matrix-Vektor-Multiplikation

Ein Bild, das Text, Schrift, weiß, Typografie enthält.

Automatisch generierte Beschreibung

Dyadisches (äusseres) Produkt

Ein Bild, das Text, Klebezettel enthält.

Automatisch generierte Beschreibung

Kronecker-Produkt

Ein Bild, das Text, Schrift, Handschrift, Kalligrafie enthält.

Automatisch generierte BeschreibungKhatri-Rao-Produkt

Ein Bild, das Text, Schrift, Handschrift, weiß enthält.

Automatisch generierte BeschreibungHadamard-Produkt

Ein Bild, das Diagramm, Text, Reihe, Schrift enthält.

Automatisch generierte Beschreibung