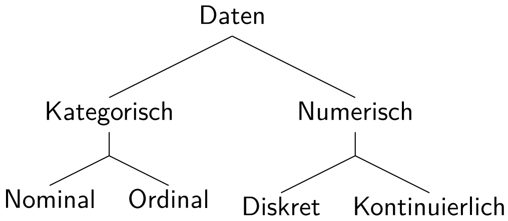
Von Daten zu Vorhersagen

Tim Barz-Cech

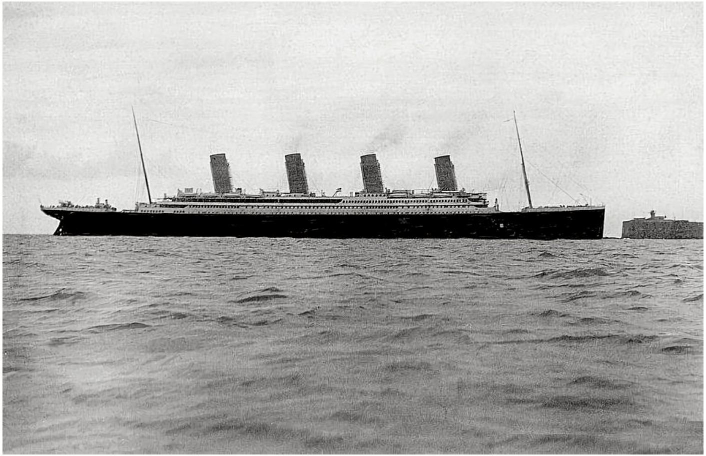
Einführung in das Machine Learning: Woche 3

Technische Hochschule Lübeck 15.09.2025

Was sind Daten? (1/2): Wiederholung



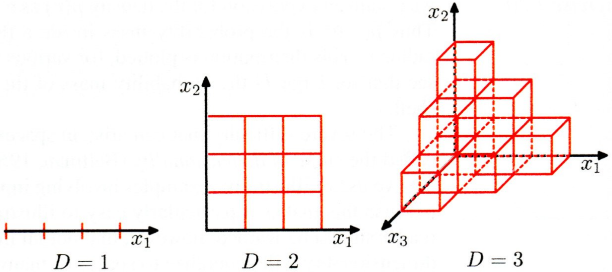
Was sind Daten? (2/2): Der Titanic-Datensatz



Die Titanic ist ein berühmtes Schiff, welches auf seiner Jungfernfahrt einen Eisberg rammte und sank

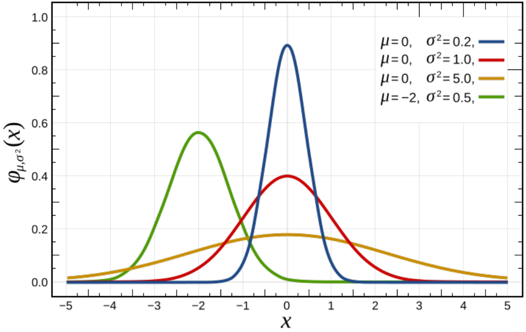
Quelle (klicken)

Der Curse of Dimensionality



Mit höherer Anzahl von Dimensionen, erhöht sich die Anzahl der Einheits-Hyperwürfels, die benötigt werden, um einen Hyperwürfel mit Seitenlänge der doppelten Einheit zu füllen [Bishop, 2006, S. 35 Abb. 1.21].

Missing Values (1/3): Mean



Verschiedene Normalverteilungen. Sofern ausreichend Samples vorliegen nähert sich der Durchschnitt dem wahren Mean (μ) Quelle (klicken)

In der Praxis ist der wahre, stochastische Mean meist unbekannt, daher ermitteln wir einen Annäherungswert durch den Durchschnitt bzw. empirischen Mean (vgl. Graphik links).

Sei $D = \{d_i | i \in \mathbb{N}\}$ ein Datensatz mit den Datenpunkten d_i, dann ist der empirische Mean oder der Durchschnitt: $Mean := \frac{1}{|D|} \sum_{i=1}^{n} d_i$

Missing Values (2/3): Median

1, 3, 3, **6**, 7, 8, 9

Median = $\underline{\underline{6}}$

1, 2, 3, **4**, **5**, 6, 8, 9

Median = $(4 + 5) \div 2$

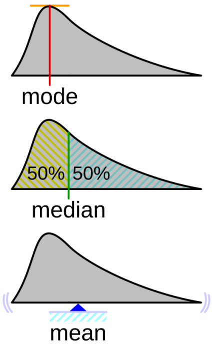
= $\underline{\underline{4.5}}$

Der Median ist der *mittlere Wert* einer geordneten Sequenz und damit streng vom Mittelwert (letzte Folie) zu unterscheiden. Der Mittelwert der dargestellten Sequenz wäre im Gegensatz zum dargestellten Median: $Mean = \frac{1+3+3+6+7+8+9}{7} = \frac{37}{7} \neq 6$ Quelle (klicken)

Missing Values (3/3): Mode & Vergleich

Der Mode ist das am häufigsten auftretene Element eines Datensatzes.

Sei $D = \{d_i | i \in \mathbb{N}\}$ ein Datensatz mit den Datenpunkten d_i, dann ist:
$Mode := \text{argmax}_{d_i \in D} d_i$



Je für verschiedene Datenarten anwendbar: Mean kontinuierliche (und selten diskrete) Daten, Median ordinal, diskrete und kontinuierliche Daten, Mode alle Datenarten Quelle (klicken)

Evaluation (1/2): Die Confusion Matrix

	True Class	
Predicted Class	Positive (c_1)	Negative (c_2)
Positive (c_1)	True Positive (TP)	False Positive (FP)
Negative (c_2)	False Negative (FN)	True Negative (TN)

Eine Übersichtsdarstellung der Confusion Matrix [Zaki and Wagner, 2014, S. 553 Tab. 22.2]

Sei $D = \{(d_i, y_i^{true}, y_i^{pred}) | i \in \mathbb{N}\}$ ein gelabelter Datensatz mit den Datenpunkten d_i, den wahren Labeln y_i^{true} und den prädiktierten Labeln y_i^{pred}. Gegeben den Bezeichnern von oben, dann ist für ein binäre Klassifikation mit positiver Klasse c_1 und negativer Klasse c_2:

- $TP := |\{d_i | y_i^{pred} = y_i^{true} = c_1\}|$
- $FP := |\{d_i | y_i^{pred} = c_1 \wedge y_i^{true} = c_2\}|$
- $FN := |\{d_i | y_i^{pred} = c_2 \wedge y_i^{true} = c_1\}|$
- $TN := |\{d_i | y_i^{pred} = y_i^{true} = c_2\}|$

Evaluation (2/2): Metriken

Betrachten wir eine binäre Klassifikation, sei $D = \{(d_i, y_i^{true}, y_i^{pred}) | i \in \mathbb{N}\}$ ein gelabelter Datensatz mit den Datenpunkten d_i, den wahren Labeln y_i^{true} und den prädiktierten Labeln y_i^{pred}, seien TP, TN, FP, FN definiert wie in der vorigen Folie, dann sind (für die positive Klasse c_1):

- $Accuracy_{c_1} := \frac{TP+TN}{|D|}$
- $Precision_{c_1} := \frac{TP}{TP+FP}$
- $Recall_{c_1} := \frac{TP}{TP+FN}$
- Analog für die negative Klasse c_2 (jedoch $Accuracy_{c_1} = Accuracy_{c_2} = Accuracy$)

Zusammenfassung

- Versuchen Sie die Anzahl der Dimensionen wann immer mögliche zu verringern (Curse of Dimensionality).
- Der hochdimensionale Raum ist erstaunlich leer und daher schwer zu klassifizieren [Verleysen and François, 2005].
- Der Umgang mit Missing Values ist von Datenart und Verteilung im Datensatz abhängig. Wählen Sie vorsichtig und weise.
- Standardmetriken können dazu benutzt werden die Qualität eines Datensatzes einzuschätzen.

Weiterführende Literatur I

[Adam et al., 2019] Adam, S. P., Alexandropoulos, S.-A. N., Pardalos, P. M., and Vrahatis, M. N. (2019). No free lunch theorem: A review. In *Approximation and Optimization: Algorithms, Complexity and Applications*, pages 57–82. Springer.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, 1 edition.

[Cech et al., 2025] Cech, T., Wegen, O., Atzberger, D., Richter, R., Scheibel, W., and Döllner, J. (2025). Standardness clouds meaning: A position regarding the informed usage of standard datasets.

[Verleysen and François, 2005] Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired systems: 8th International Work-Conference on Artificial Neural Networks*, IWANN '05, pages 758–770. Springer.

[Zaki and Wagner, 2014] Zaki, M. J. and Wagner, M. J. (2014). *Data Mining and Analysis: Fundamental Concepts And Algorithms*. Cambridge University Press, 1 edition.