Machine Learning (Online) Übung 1

Das Ziel dieser Übung ist, den Preis einer Immobilie mit Hilfe von linearen Regressionen aus verschiedenen Attributen abzuschätzen.

Aufgaben:

1) **(2 Punkte)** Laden und visualisieren Sie den Datensatz house_data.csv.

2) **(6 Punkte)** Schätzen Sie 'price' mit einer linearen Regression aus dem Attribut 'sqft_living' mit dem im Kurs besprochenen *Gradient-Descent* Verfahren. Visualisieren Sie die Kosten per Iteration um sicherzustellen, dass das Verfahren korrekt funktioniert.

   Visualisieren Sie die Regressionsgerade zusammen mit allen Datenpunkten. Erstellen Sie weiter einen Tukey-Anscombe-Plot und interpretieren Sie diesen.

   ➔ http://stat.ethz.ch/~stahel/courses/regression/reg-resanal.pdf

3) **(4 Punkte)** Wiederholen Sie Aufgabe 2) mit der transformierten Output-Variable *log*('price'). Erstellen Sie zusätzlich ein Histogramm der Residuen und interpretieren Sie dieses bezüglich Standardabweichung.

4) **(4 Punkte)** Berechnen sie den *mean absolute percentage error* (MAPE) für die Modelle aus Aufgabe 2 und 3. Visualisieren Sie die Verteilung des *percentage error*'s der beiden Modelle mittels Histogrammen und diskutieren Sie dieses.

5) **(2 Punkte)** Visualisieren Sie mit einer geeigneten Farbskalierung die Output-Variablen *log*('price') als Punkte in den geografischen Koordinaten 'long'/'lat' (geografische Länge/Breite) und diskutieren Sie das Ergebnis.

6) **(2 Punkte)** Wiederholen Sie Aufgabe 5) mit dem Attribut 'zipcode' anstelle von *log*('price') und interpretieren Sie den Einfluss von 'zipcode' auf den Immobilienpreis.

7) **(4 Punkte)** Feature Engineering: Codieren Sie das kategorische Attribut 'zipcode' mittels *one-hot encoding* und schätzen Sie *log*('price') mit einer neuen linearen Regression aus 'sqft_living' und den neu kreierten Attributen. Erstellen Sie ein Histogramm der Residuen und diskutieren Sie diese. Wie verhält sich der Mittelwert, bzw. die Standardabweichung im Vergleich zu Aufgabe 3)?

8) **(3 Punkte)** Wiederholen Sie die Aufgabe 7) mit den vier zusätzlichen Attributen ['bedrooms', 'bathrooms', 'grade', 'yr_built'] und diskutieren Sie den Unterschied zu 7).

9) **(4 Punkte)** Was müssten Sie anpassen wenn Sie den MAPE als Kostenfunktion für *Gradient-Descent* verwenden möchten? Was könnte passieren, wenn Sie den RSS als Kostenfunktion verwenden, jedoch am Ende das Modell mittels MAPE evaluieren?

Abgabe: Beschreiben Sie ihre Lösung in einem PDF und senden Sie dieses zusammen mit dem Code als Zip-Datei in der Form

Nachname_Vorname.zip

bis spätestens 2. April 2018 an lukas.neukom@fhnw.ch.

Ein paar Vorgaben für die Abgabe:

- verwenden Sie die im Kurs entwickelte oder eine eigene Implementation für die lineare Regression; nicht die *built-in* Lineare Regression von MATLAB.
- der Code muss ohne Änderungen unsererseits lauffähig sein
- Plots müssen beschriftet sein (Achsen, Titel, kurze Beschreibung des Plots)
- eine Intepretation eines Plots beinhaltet eine Beschreibung sowie eine Diskussion der gezeigten Daten

Beispiel einer *one-hot* Codierung:

id	Color
1	White
2	Red
3	Black
4	Purple
5	Gold

id	White	Red	Black	Purple	Gold
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1