# 1. Grundlagen

- Sortiere folgende Modelle nach Interpretierbarkeit: [Decision Tree, SVM, Lineares Modell, Neuronales Netz].
- Qualitätseigenschaften einer Erläuterung: Welche sind am wenigsten hilfreich für die Akzeptanz von Menschen? Erkläre!
- Nenne zwei gegensätzliche Aspekte menschenfreundlicher Erklärungen. Erkläre!
- Was ist eine modellagnostische Erklärungsmethode?
- Erkläre lokale und globale Erklärmethoden anhand des 'ImageNet'-Klassifikators.

# 2. Lokal Agnostische Erklärmethoden

- Erkläre den LIME-Algorithmus für ein Blackbox-Modell.
- Nimm an, es gibt ein Surrogatmodell mit einem (1) kontinuierlichen Merkmal. Wie viele Parameter besitzt das lineare Modell?
  - ↪ ohne Diskretisierung
  - ↪ mit Diskretisierung (4 Intervalle)
- Was ist der Vorteil der Diskretisierung innerhalb des LIME-Algorithmus?
- Nenne zwei Bedingungen von sinnvollen Counterfactuals.
- Berechne den Shapley-Wert für Spieler 2.

| | | | | |
|---|---|---|---|---|
| 1, 2, 3 | S = ∅<br>0 | S = {1}<br>+1000 | S = {1, 2}<br>+4000 | S = {1, 2, 3}<br>+6000 |
| 1, 3, 2 | S = ∅<br>0 | S = {1}<br>+1000 | S = {1, 3}<br>+4000 | S = {1, 2, 3}<br>+6000 |
| 2, 1, 3 | S = ∅<br>0 | S = {2}<br>+2000 | S = {1, 2}<br>+4000 | S = {1, 2, 3}<br>+6000 |
| 2, 3, 1 | S = ∅<br>0 | S = {2}<br>+2000 | S = {2, 3}<br>+3000 | S = {1, 2, 3}<br>+6000 |
| 3, 1, 2 | S = ∅<br>0 | S = {3}<br>+3000 | S = {1, 3}<br>+4000 | S = {1, 2, 3}<br>+6000 |
| 3, 2, 1 | S = ∅<br>0 | S = {3}<br>+3000 | S = {2, 3}<br>+3000 | S = {1, 2, 3}<br>+6000 |

- Erkläre den Zusammenhang zwischen $E[f(x)]$, den SHAP-Werten und der Vorhersage $f(x)$.

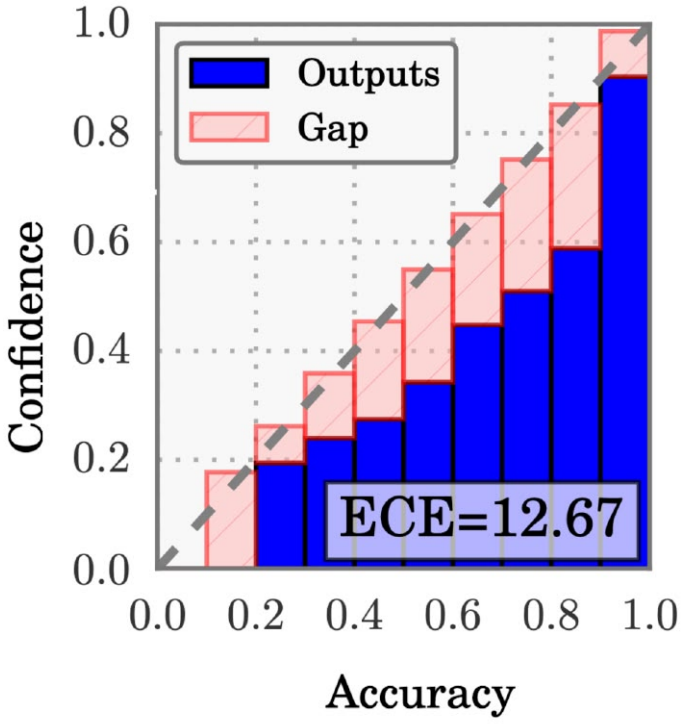
- Wähle ein Feature aus dem gegebenen SHAP-Plot aus und erkläre dessen Bedeutung! (Hier einfach einen x-beliebigen SHAP-Plot vorstellen, in der Prüfung war es der Fahrrad-Datensatz aus der Vorlesung.)
- Sind Richtung und Größe der Werte im Plot für die Ausgabe (Anzahl Fahrräder) plausibel?

# 3. Global Agnostische Erklärmethoden

- Erkläre den PFI-Algorithmus (Permutation Feature Importance)!
- Welcher Zusammenhang besteht zwischen CP, ICE und PD?
- Zeichne basierend auf dem dargestellten PD-Plot passende ICE-Kurven ein. (Hier auch einfach einen x-beliebigen PD-Plot ausdenken (logischerweise nicht mit ICE-Plot überlagert).)

# 4. Uncertainty und Kalibrierung

- Ist epistemische Unsicherheit verringerbar? Wenn ja, wie?
- Nenne ein Beispiel oder eine Quelle für heteroskedastische aleatorische Unsicherheit bei der Klassifikation von MNIST und ImageNet.
- Beschreibe eine beliebige Ensembling-Methode und erkläre, wie sie bei der Ermittlung der Unsicherheit hilft.
- Schätze anhand des dargestellten Plots (Confidence vs. Accuracy), wie gut das Modell kalibriert ist.

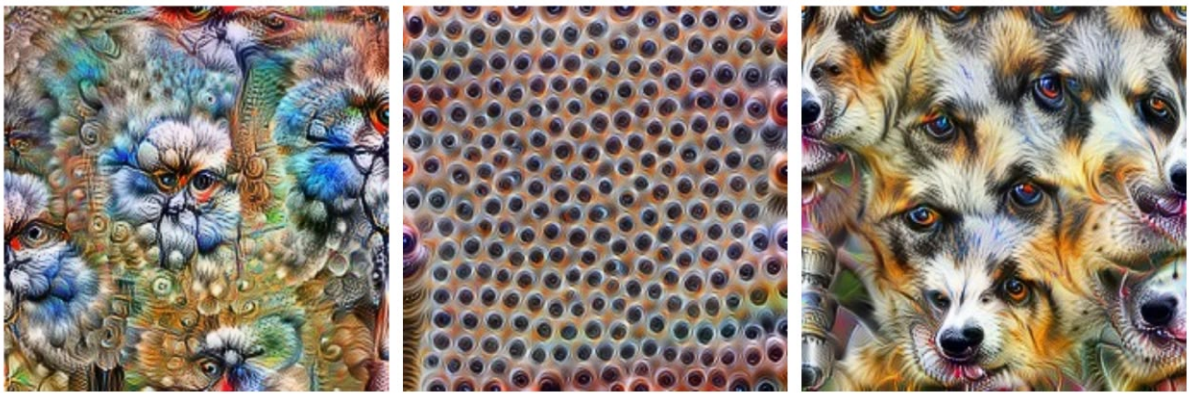


– An welcher Stelle wird der Parameter $T$ für Temperature Scaling in der Softmax-Funktion eingesetzt? Gib die komplette Formel an.

– Ein Modell ist überkonfident bei $T = 1$. In welche Richtung muss $T$ angepasst werden, um das Modell besser zu kalibrieren?

# 5. Feature Visualisation u. Saliency Maps

– Bringe die gegebenen, durch Feature Visualization entstandenen Bilder aus dem 'GoogleNet'-Modell in die Reihenfolge, wie sie im Modell vorkommen würden.



– Erkläre das Prinzip der Network Dissection.

– Nenne zwei Voraussetzungen für die Anwendung von Grad-CAM.

– Gegeben sei die CAM-Formel. Wie wird $\alpha_k^c$ für Grad-CAM berechnet? Nenne einen Nachteil von Grad-CAM.

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

# 6. Repräsentationsanalyse

– Wie lässt sich überprüfen, ob ein berechneter TCAV-Score nicht rein zufällig ist?

– Gegeben sei ein Vision Transformer mit CLS-Token. Lassen sich 'ImageNet'-Klassen anhand ihrer Repräsentationen voneinander trennen? *Hinweis: Die 'ImageNet'-Klassen sind sehr wahrscheinlich* ***nicht*** *linear separierbar.*

  ↪ Wähle eine dafür geeignete Methode zur Dimensionsreduktion und erkläre diese!

  ↪ Ist es sinnvoll, zur Extraktion den CLS-Token aus der ersten Schicht zu nutzen? Erkläre!

  ↪ Schlage eine andere Extraktionsmethode vor. Warum ist diese besser?

## 7. XAI für Transformer

- Nenne eine Gemeinsamkeit und einen Unterschied von Attention Flow und Attention Rollout.
- Welche Dimension hat das Ergebnis von Attention Rollout bei einem Vision Transformer, der das Originalbild in 14x14 Patches teilt?
- Erkläre die Rolle der Einheitsmatrix $I$ beim Attention Rollout.
- Was ist eine typische Anpassung des Attention Rollout, um die resultierenden Saliency Maps besser interpretierbar zu machen?

## 8. KI Sicherheit

- Gegeben sind 3 Angriffe. Welcher davon stellt eine Adversarial Attack dar? Erkläre!

1.


2.


3.


- Wähle eine Risikoklasse des AI Acts der EU aus, beschreibe diese und nenne die Auflagen für diese Stufe.
- Was versteht man unter Prompt Injection? Schlage eine Guardrail-Maßnahme dagegen vor.
- Nenne 2 Gefahren von Reinforcement Learning from Human Feedback (RLHF).