CS480/680: Introduction to Machine Learning

Lec 20: Model Interpretability

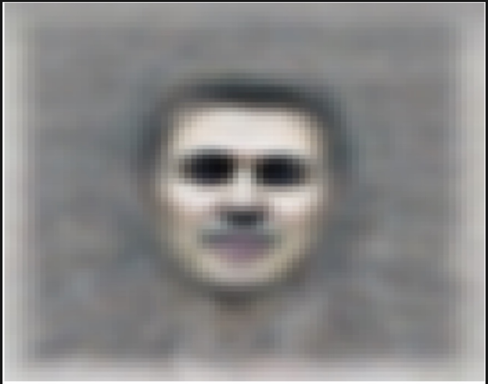
Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

April 4, 2024

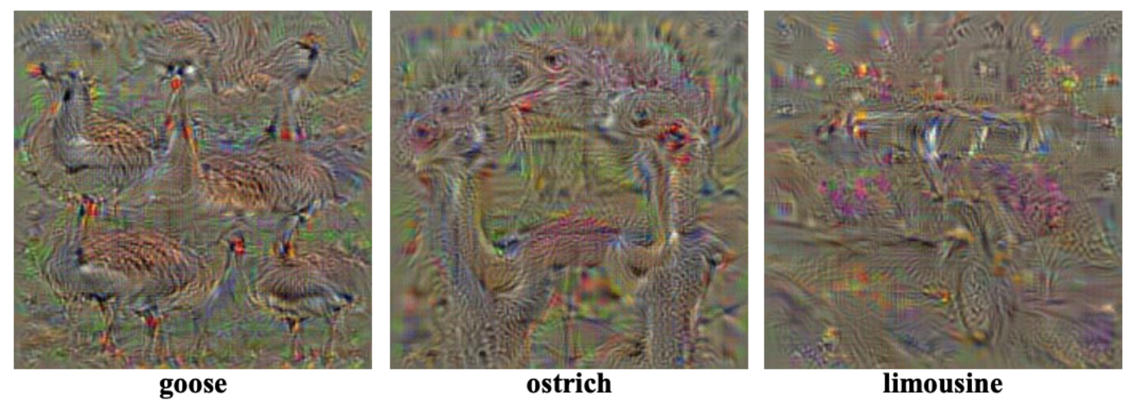
Activation Maximization

- To understand a neuron activation, fix the network weights
- Enumerate test set or run (projected) grad ascent on input



Q. V. Le et al. "Building high-level features using large scale unsupervised learning". In: *Proceedings of the 29th International Conference on Machine Learning*. 2012.

Gradient Saliency



K. Simonyan et al. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *ICLR workshop*. 2017, R. R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

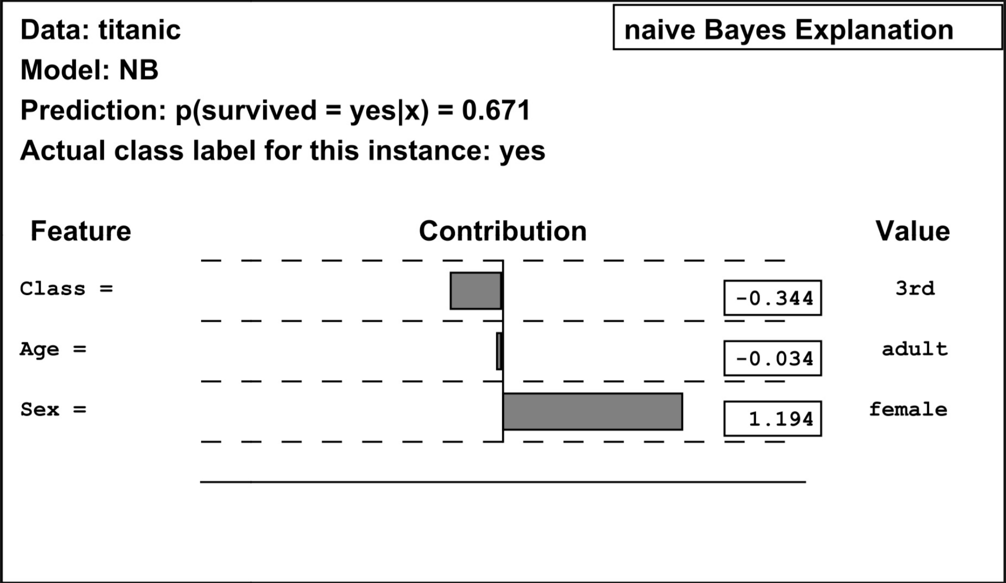
OR Example

$$y = x_1 \texttt{ or } x_2$$

- $n = 2$; consider $x_1 = x_2 = 1$
- Gradient methods do not work
  - fix x_1, conclude that x_2 does not matter
  - fix x_2, conclude that x_1 does not matter
  - conclude neither x_1 or x_2 matters ...
- $u(1) = u(2) = u(1, 2) = 1$ and 0 else
- Banzhaf value: $p_s \equiv \frac{1}{2^{n-1}} \implies \phi_1 = \phi_2 = \frac{1}{2}$
- Shapley value: $p_s = \frac{s!(n-s-1)!}{n!} \implies \phi_1 = \phi_2 = \frac{1}{2}$

Coalition Game

- n is the number of "players"
- $u : 2^{[n]} \to \mathbb{R}$ the "payoff" function
  - w.l.o.g. $u(\emptyset) = 0$, where $\emptyset$ is the baseline
- Examples:
  - each feature is a player (feature valuation)
  - each training example is a player (data valuation)
  - each neuron is a player
  - performance metric (e.g. accuracy) is payoff
- Valuation: what is the value of each player i?

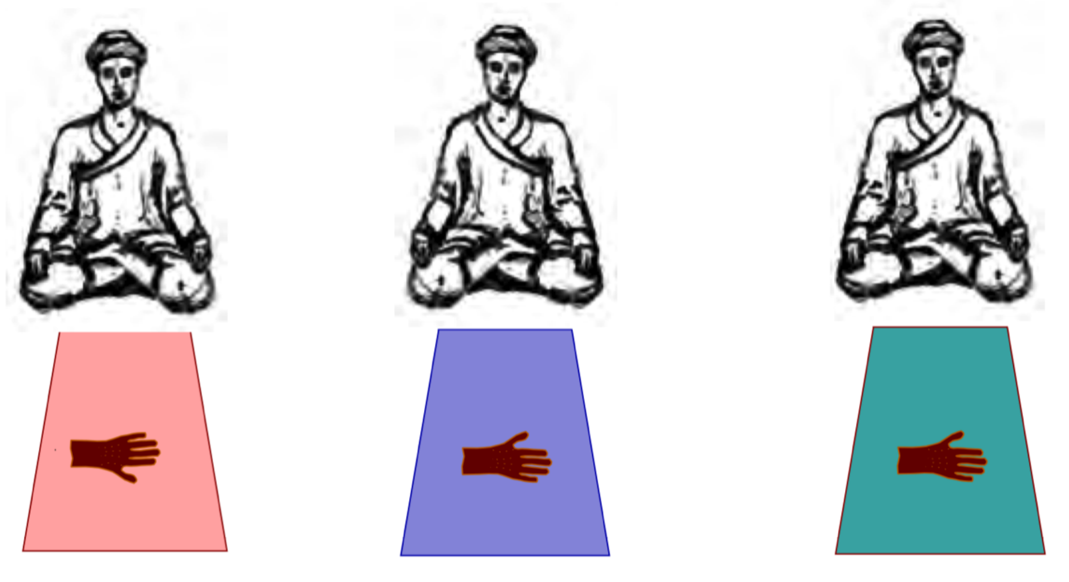


E. Štrumbelj and I. Kononenko. "An Efficient Explanation of Individual Classifications using Game Theory". *Journal of Machine Learning Research*, vol. 11 (2010), pp. 1–18.

Probabilistic Value

- Given a set function $u : 2^{[n]} \to \mathbb{R}$, e.g., accuracy trained on subset of data
- Find an additive approximation $\phi : 2^{[n]} \to \mathbb{R}$, where $\phi(S) = \sum_{i \in S} \phi(\{i\})$
- Marginal contribution of i: $u(S \cup \{i\}) - u(S \setminus \{i\})$
- Leave-one out: $u([n]) - u([n] \setminus \{i\})$
- (Symmetric) probabilistic value: $\phi_i^p = \phi^p(\{i\}) = \sum_{S \not\ni i} p_s \cdot [u(S \cup \{i\}) - u(S)]$
  - from now on, $s = |S|$

R. J. Weber. "Probabilistic values for games". In: *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Ed. by A. E. Roth. Cambridge University Press, 1988, pp. 101–120.



A. R. Karlin and Y. Peres. "Game Theory, Alive!" American Mathematical Society, 2017.

$$\boxed{\phi_i = \sum_{S \not\ni i} p_s \cdot [u(S \cup \{i\}) - u(S)]}$$

- $n = 3$
- $u(1,2) = u(1,3) = u(1,2,3) = 1$ and 0 else
- Banzhaf value: $p_s \equiv \frac{1}{2^{n-1}} \implies \phi_1 = \frac{1}{4} + \frac{1}{4} + \frac{1}{4}, \quad \phi_2 = \phi_3 = \frac{1}{4}$
- Shapley value: $p_s = \frac{s!(n-s-1)!}{n!} \implies \phi_1 = \frac{1}{6} + \frac{1}{6} + \frac{1}{3}, \quad \phi_2 = \phi_3 = \frac{1}{6}$

J. F. Banzhaf III. "Weighted Voting Doesn't Work: A Mathematical Analysis". *Rutgers Law Review*, vol. 19 (1965), pp. 317–343.
L. S. Shapley. "A Value for n-person Games". In: *Contributions to the Theory of Games*. Vol. 2. 1953, pp. 307–318.

John Francis Banzhaf III

Lloyd Stowell Shapley, Nobel Prize (2012)

One Man, One Vote?

State Name (1)	Population 1960 Census	Electoral Vote 1964	Relative Voting Power (2)	Percent Excess Voting Power (3)	Percent Deviation From Average Voting Power (4)
Alabama	3266740.	10	1.632	63.2	–3.0
Alaska	226167.	3	1.838	83.8	9.2
Arizona	1302161.	5	1.281	28.1	–23.9
Arkansas	1786272.	6	1.315	31.5	–21.9
California	15717204.	40	3.162	216.2	87.9
Colorado	1753947.	6	1.327	32.7	–21.1
Connecticut	2535234.	8	1.477	47.7	–12.2
Delaware	446292.	3	1.308	30.8	–22.3
Dist. of Columbia	763956.	3	1.000	.0	–40.6
Florida	4951560.	14	1.870	87.0	11.1
Georgia	3943116.	12	1.789	78.9	6.3
Hawaii	632772.	4	1.468	46.8	–12.8
Idaho	667191.	4	1.429	42.9	–15.1
Illinois	10081158.	26	2.491	149.1	48.0
Indiana	4662498.	13	1.786	78.6	6.1
Iowa	2757537.	9	1.596	59.6	–5.2
Kansas	2178611.	7	1.392	39.2	–17.3
Kentucky	3038156.	9	1.521	52.1	–9.6
Louisiana	3257022.	10	1.635	63.5	–2.9
Maine	969265.	4	1.186	18.6	–29.5
Maryland	3100689.	10	1.675	67.5	–.4
Massachusetts	5148578.	14	1.834	83.4	9.0
Michigan	7823194.	21	2.262	126.2	34.4
Minnesota	3413864.	10	1.597	59.7	–5.1
Mississippi	2178141.	7	1.392	39.2	–17.3
Missouri	4319813.	12	1.710	71.0	1.6
Montana	674767.	4	1.421	42.1	–15.5
Nebraska	1411330.	5	1.231	23.1	–26.9
Nevada	285278.	3	1.636	63.6	–2.8
New Hampshire	606921.	4	1.499	49.9	–10.9
New Jersey	6066782.	17	2.063	106.3	22.6
New Mexico	951023.	4	1.197	19.7	–28.9
New York	16782304.	43	3.312	231.2	96.8
North Carolina	4556155.	13	1.807	80.7	7.4
North Dakota	632446.	4	1.468	46.8	–12.8
Ohio	9706397.	26	2.539	153.9	50.9
Oklahoma	2328284.	8	1.541	54.1	–8.4
Oregon	1768687.	6	1.321	32.1	–21.5
Pennsylvania	11319366.	29	2.638	163.8	56.8
Rhode Island	859488.	4	1.259	25.9	–25.2
South Carolina	2382594.	8	1.524	52.4	–9.5
South Dakota	680514.	4	1.415	41.5	–15.9
Tennessee	3567089.	11	1.721	72.1	2.3
Texas	9579677.	25	2.452	145.2	45.7
Utah	890627.	4	1.237	23.7	–26.5
Vermont	389881.	3	1.400	40.0	–16.8
Virginia	3966949.	12	1.784	78.4	6.0
Washington	2853214.	9	1.569	56.9	–6.8
West Virginia	1860421.	7	1.506	50.6	–10.5
Wisconsin	3951777.	12	1.788	78.8	6.2
Wyoming	330066.	3	1.521	52.1	–9.6

J. F. Banzhaf III. "One Man, 3.312 Votes: A Mathematical Analysis of the Electoral College". *Villanova Law Review*, vol. 13, no. 2 (1968), pp. 304–332.

Random Order Value

- Let π be a permutation of $[n] := \{1, 2, \ldots, n\}$
- Suppose $i = \pi(k)$ and define

$$\psi_i(u, \pi) = u[\underbrace{\pi(1), \ldots, \pi(k)}_{\text{when } i \text{ joins}}] - u[\underbrace{\pi(1), \ldots, \pi(k-1)}_{\text{before } i \text{ joins}}]$$

- Randomize over permutations: $\boxed{\phi_i(u) = \mathbb{E}_\pi \psi_i(u, \pi)}$
- What happens if we sum all values?

$$\sum_i \phi_i(u) =?$$

R. J. Weber. "Probabilistic values for games". In: *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Ed. by A. E. Roth. Cambridge University Press, 1988, pp. 101–120.

Shapley's Axioms

- Linear: $\phi_i(u+v) = \phi_i(u) + \phi_i(v)$
- Symmetry: if $u(S \cup i) = u(S \cup j)$ for all S with $i, j \notin S$, then $\phi_i = \phi_j$
- Null: if $u(S \cup i) = u(S)$ for all S with $i \notin S$, then $\phi_i = 0$
- Efficient: $\sum_i \phi_i = u([n])$

$$\phi_i(u) = \sum_{S \not\ni i} \frac{s!(n-s-1)!}{n!} \cdot [u(S \cup \{i\}) - u(S)], \qquad \pi \sim \text{Uniform}$$

L. S. Shapley. "A Value for n-person Games". In: *Contributions to the Theory of Games*. Vol. 2. 1953, pp. 307–318.

How to Estimate Probabilistic Value?

$$\phi_i^p = \phi^p(\{i\}) = \sum_{S \not\ni i} p_s \cdot [u(S \cup i) - u(S)]$$

Algorithm 1: Monte Carlo estimation of probabilistic value

Input: utility u, probability p

for $i = 1, \ldots, n$ **do**
  $\varphi_i \leftarrow 0$
  for $k = 1, \ldots, m$ **do**
    sample a random subset $S \not\ni i$ with probability $\propto \binom{n-1}{s} p_s$
    $\varphi_i \leftarrow \varphi_i + [u(S \cup \{i\}) - u(S)]$ `// 2 evals of utility`
  $\hat{\phi}_i \leftarrow \varphi_i / m$

X. Deng and C. H. Papadimitriou. "On the Complexity of Cooperative Solution Concepts". *Mathematics of Operations Research*, vol. 19, no. 2 (1994), pp. 257–266.

- Suppose w.l.o.g. utility $u \in [0, 1]$
- $\hat{\phi}_i$ is an average over m i.i.d. samples
- From Hoeffding's inequality: $\Pr[|\hat{\phi}_i - \phi_i| \geq \epsilon] \leq 2\exp(-m\epsilon^2/2)$
- To achieve $\|\hat{\phi} - \phi\|_\infty \leq \epsilon$ with probability $1 - \delta$, need $O(\frac{n}{\epsilon^2} \log \frac{n}{\delta})$ samples
- Maximum sample reuse for the Banzhaf value: $O(\frac{n}{\epsilon^2} \log \frac{n}{\delta})$ for ℓ_2 norm
  - $\|\hat{\phi} - \phi\|_2 \leq \epsilon$ vs. $\|\hat{\phi} - \phi\|_\infty \leq \epsilon/\sqrt{n}$

J. Wang and R. Jia. "Data Banzhaf: A Robust Data Valuation Framework for Machine Learning". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. 2023.

Least-square Value

$$\min_{\phi \in \mathbb{R}^n} \sum_{S \subseteq [n]} q_s \cdot [u(S) - \phi(S)]^2 \quad \text{s.t.} \quad u([n]) = \sum_i \phi_i$$

- Take $q_s = p_s + p_{s-1}$ recovers (efficient normalization of) probabilistic value
  - $\sum_{S \not\ni i} p_s[u(S \cup \{i\}) - u(S)] = \sum_{S \ni i} p_{s-1} u(S) - \sum_{S \not\ni i} p_s u(S) = \sum_{S \ni i} [p_{s-1} + p_s] u(S) - \sum_S p_s u(S)$
  - Shapley value corresponds to $q_s = \frac{(s-1)!(n-1-s)!}{(n-1)!} \equiv \frac{1}{\binom{n-2}{s-1}}$
- Can approximate with $O(\frac{n}{\epsilon^2} \log \frac{n}{\delta})$ samples

---

A. Charnes et al. "Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations". In: *Econometrics of Planning and Efficiency*. 1988, pp. 123–133, L. M. Ruiz et al. "The Family of Least Square Values for Transferable Utility Games". *Games and Economic Behavior*, vol. 24, no. 1-2 (1998), pp. 109–130.

W. Li and Y. Yu. "Faster Approximation of Probabilistic and Distributional Values via Least Squares". In: *International Conference on Learning Representations (ICLR)*. 2024.

Shapley value $\subseteq$ Random order value $\subseteq$ Probabilistic value $\subseteq$ Least-square value