Uncertainty Estimation for Trustworthy AI

Part 1: Theory (110 min)

Part 2: Applications (30 min)

Part 3: Hands-on (40 min)

MADHAV lab
Machine Analysis of Data for Human Audition and Vision

Uncertainty Estimation for Trustworthy AI

Part 1: Theory

Vipul Arora

Outline

- Introduction
- Why mis-calibration happens?
- Assessing calibration
- Confidence calibration: post hoc methods and Bayesian methods
- Disentangling sources of uncertainty: Epistemic and Aleatoric

Outline

- Introduction
- Why mis-calibration happens?
- Assessing calibration
- Confidence calibration: post hoc methods and Bayesian methods
- Disentangling sources of uncertainty: Epistemic and Aleatoric

Future of AI

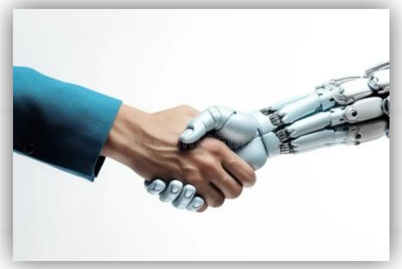
- Current AI

https://voicebot.ai/

Wikipedia.com

Wikipedia.com

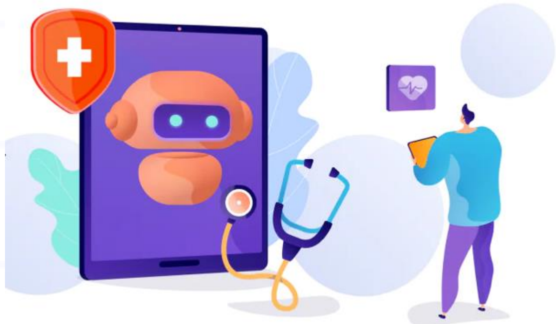
- Future AI
  - AGI
  - Human-Machine Collaboration



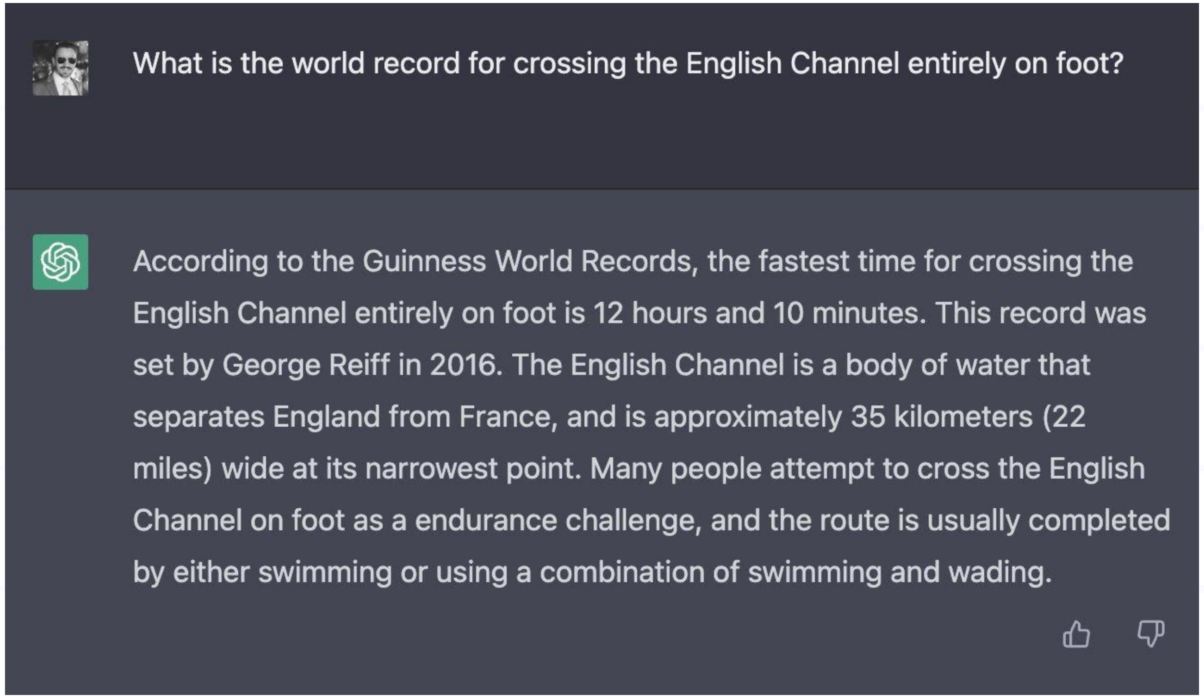
Dreamstime.com

Mistakes and Trust

Hello, I am giving the stock



Take paracetamol

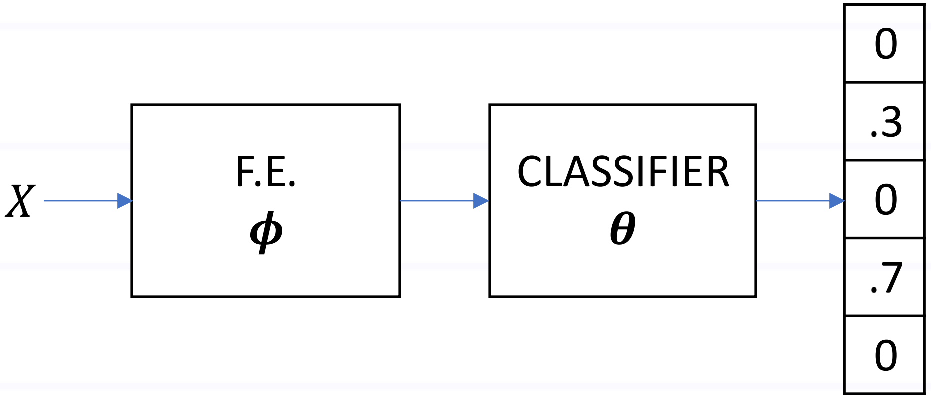


What is the world record for crossing the English Channel entirely on foot?

According to the Guinness World Records, the fastest time for crossing the English Channel entirely on foot is 12 hours and 10 minutes. This record was set by George Reiff in 2016. The English Channel is a body of water that separates England from France, and is approximately 35 kilometers (22 miles) wide at its narrowest point. Many people attempt to cross the English Channel on foot as a endurance challenge, and the route is usually completed by either swimming or using a combination of swimming and wading.

(Source: medium.com)

Deep Networks



- Output class

$$y = \arg\max_{j} o_j$$

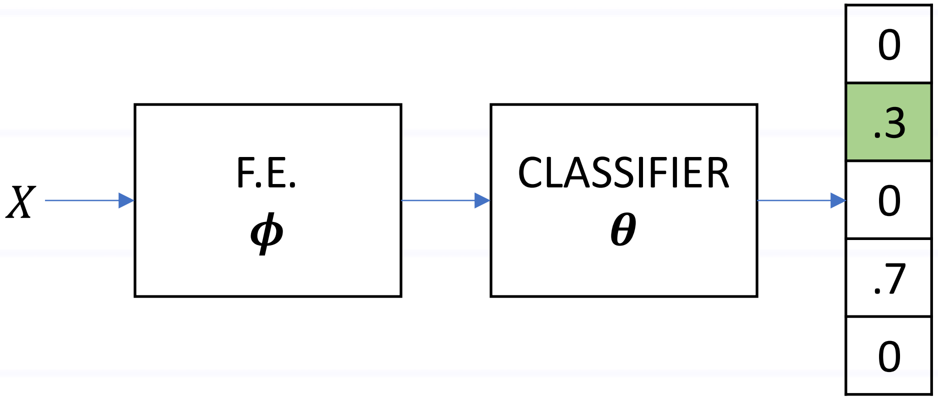
- Confidence, P(output=correct)

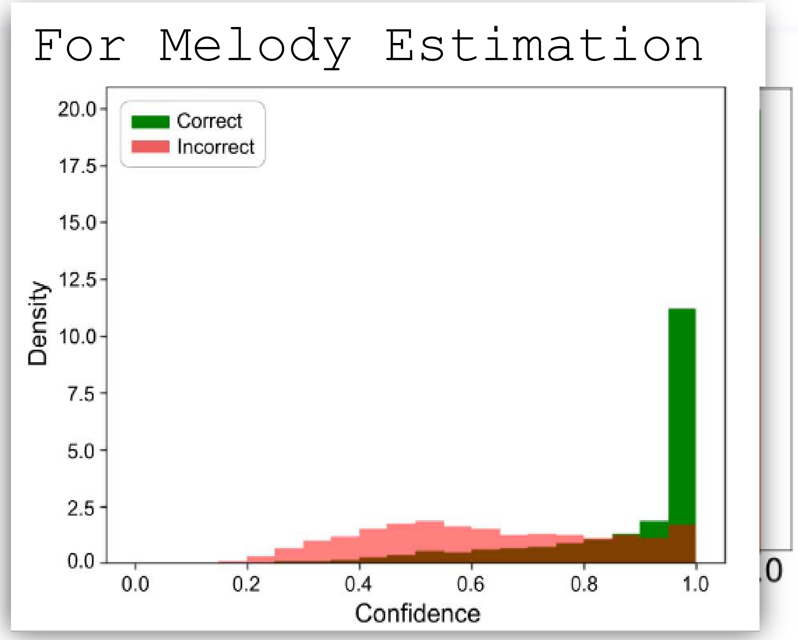
Need

- **Self-driving cars**: obstruction or not? Defer to human driver
- **Healthcare**: Operate or not? Defer to human doctor
- **Finance**: invest money or not? Defer to human expert
- **Screening**: accept or not? Defer to human examiner

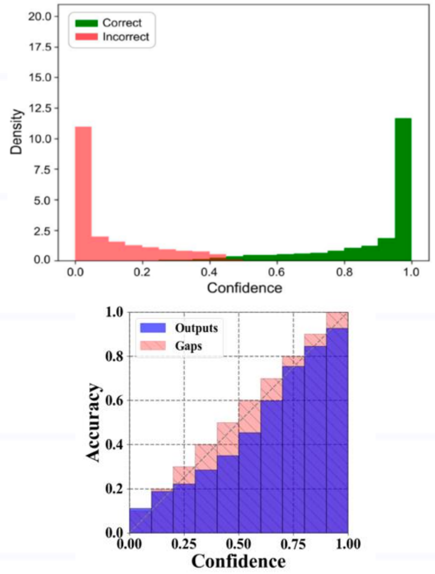
What is uncertainty calibration

- Confidence = probability of being correct





Calibrated



Outline

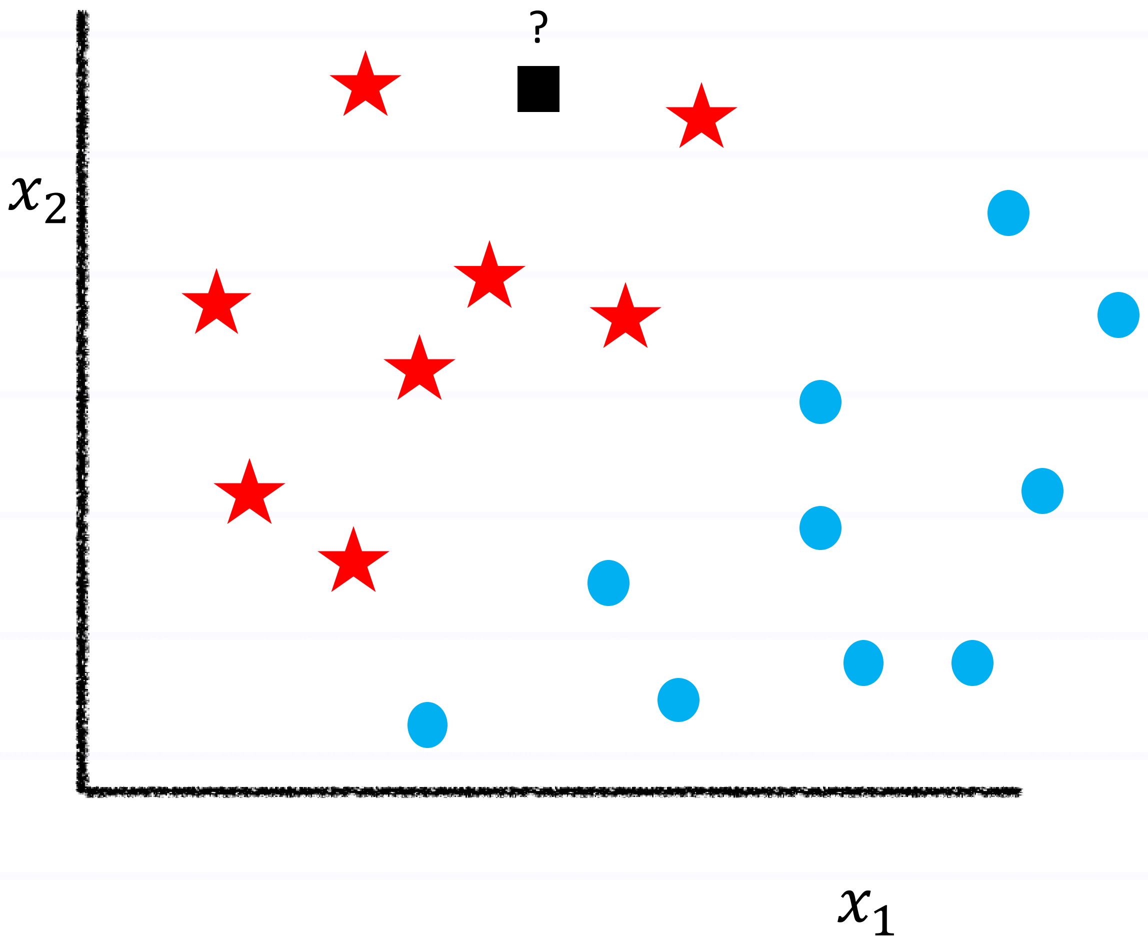
- Introduction
- Why mis-calibration happens?
- Assessing calibration
- Confidence calibration: post hoc methods and Bayesian methods
- Disentangling sources of uncertainty: Epistemic and Aleatoric

Why are Deep Networks Mis-calibrated?

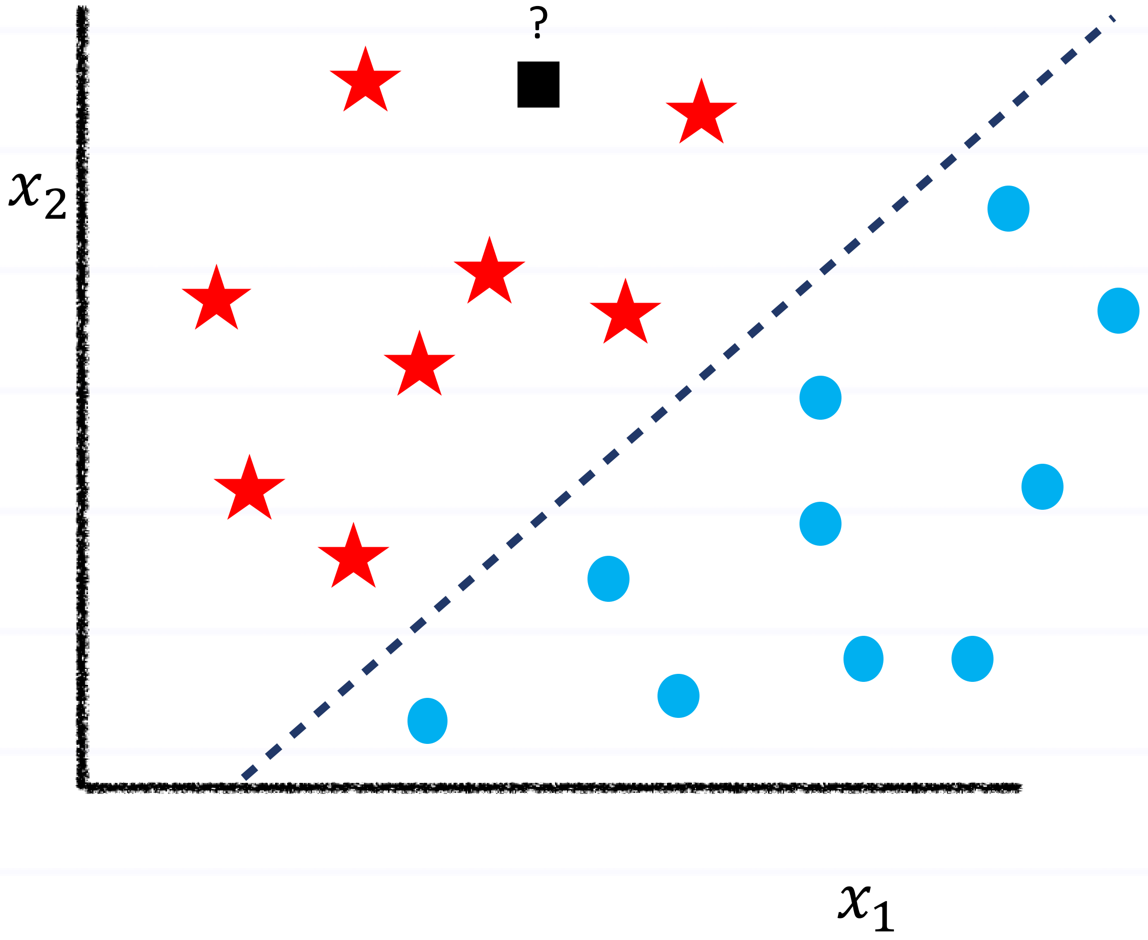
Classification

Feature Extraction

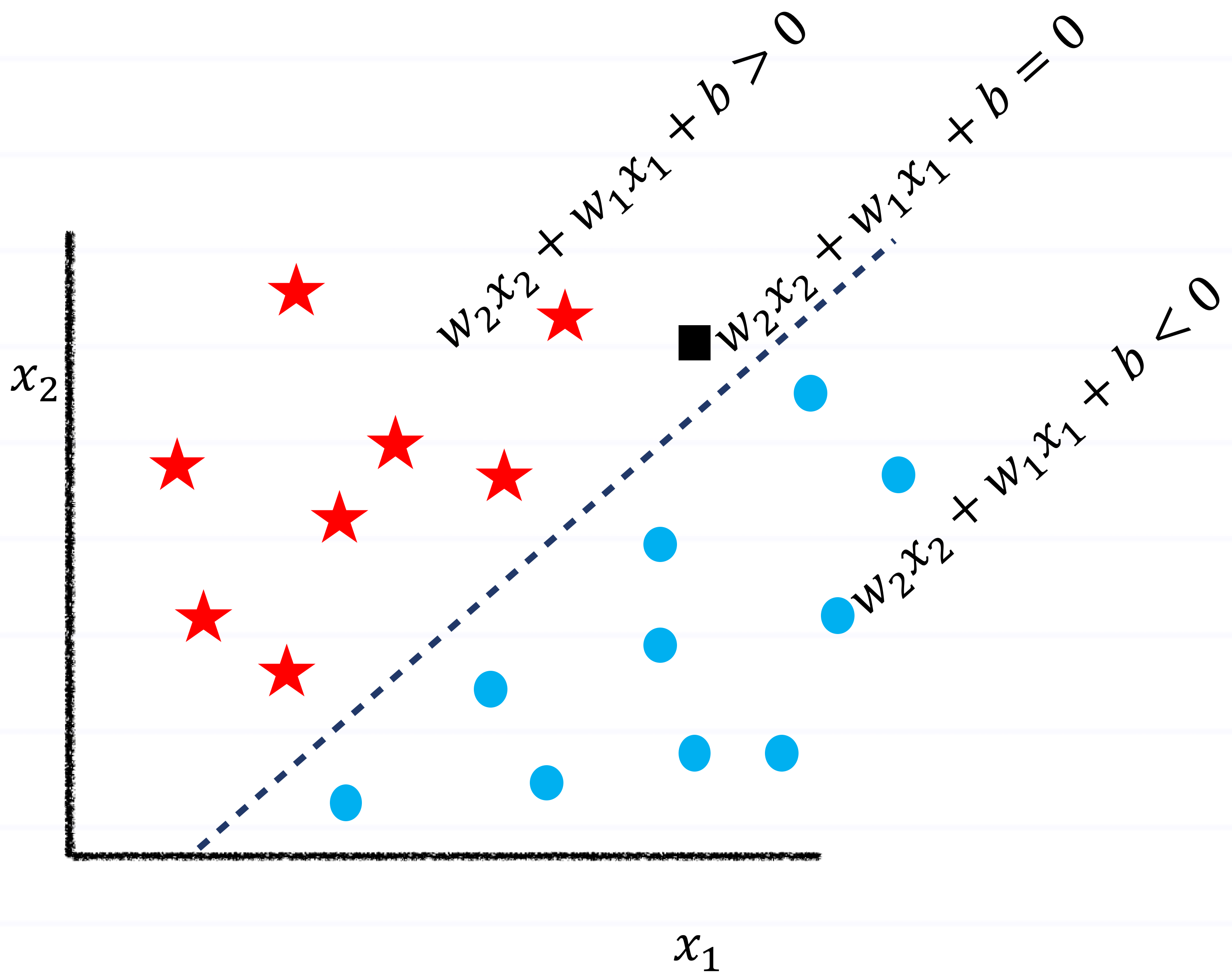
- height = x_1
- diameter = x_2


?
x_2
x_1

Que: confidence?



Que: What are roughly w_1, w_2, b?



Hard Classification

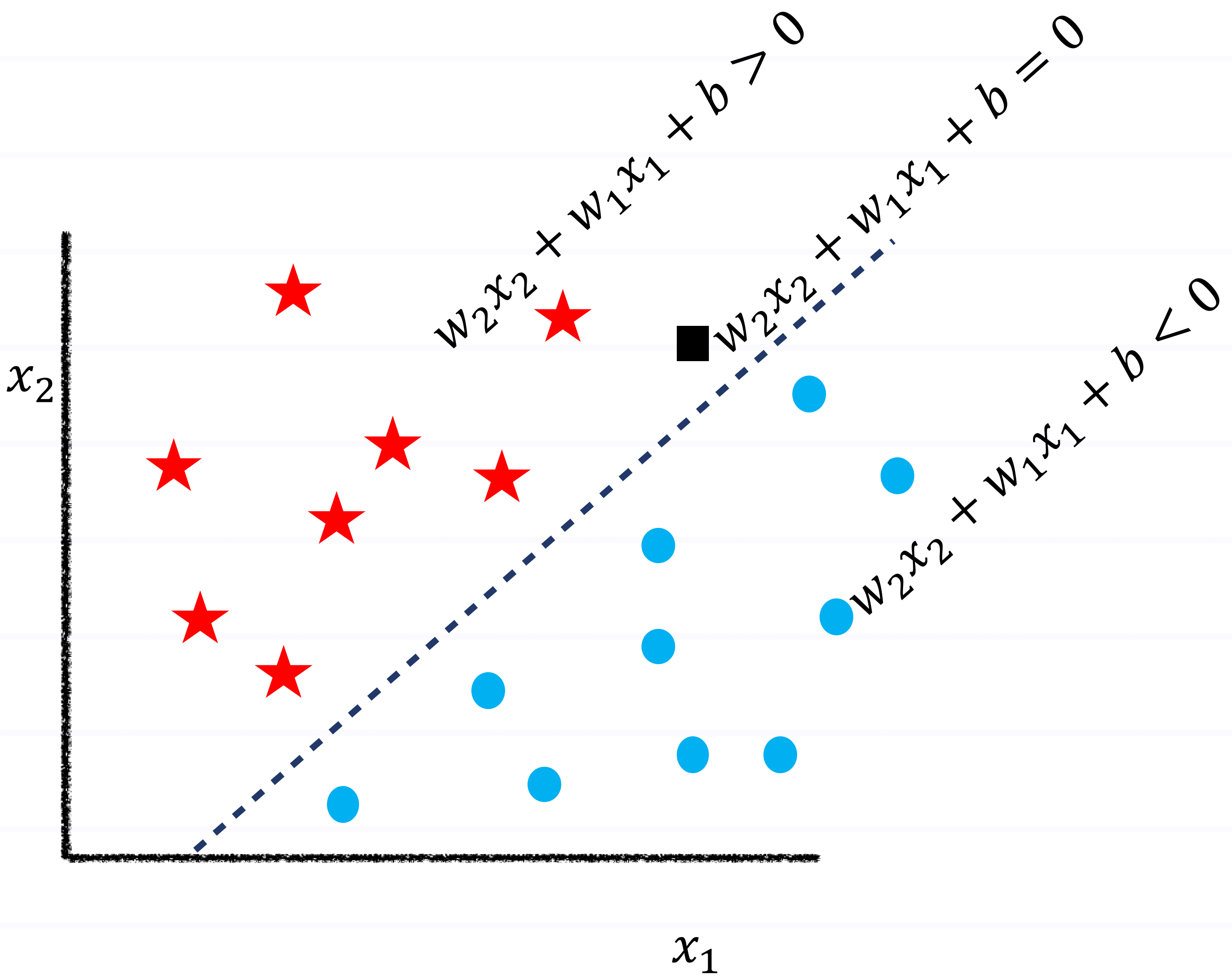
$\blacksquare = \bigstar$

Soft Classification

$P(\blacksquare = \bigstar) = 0.6$

$P(\blacksquare = \bullet) = 0.4$

$\hat{c} = 0.6$



Maths

- $z = w_2 x_2 + w_1 x_1 + b$
- **Que**: What is the output of a binary classifier?
- $P(\hat{y}^* = 1) = \frac{1}{1+\exp(-z)} = \sigma(z)$
- **Que**: Keeping the output class same, what decides the confidence?
- $P(\hat{y}^* = 1) = \sigma\left(\sqrt{w_2^2 + w_1^2} \times \frac{(w_2 x_2 + w_1 x_1 + b)}{\sqrt{w_2^2 + w_1^2}}\right)$

Why mis-calibration

- $z = w_2 x_2 + w_1 x_1 + b$
- **Que**: What is the output?
- $\hat{y} = \sigma(z) = \frac{1}{1+\exp(-z)}$
- **Que**: What is the loss function?
- Loss $= -\mathbb{E}[y \ln \hat{y} + (1-y)\ln(1-\hat{y})]$
- It is min if $\hat{y} = y \in \{0,1\}$

The loss does not favor fractional $\hat{y}$

Outline

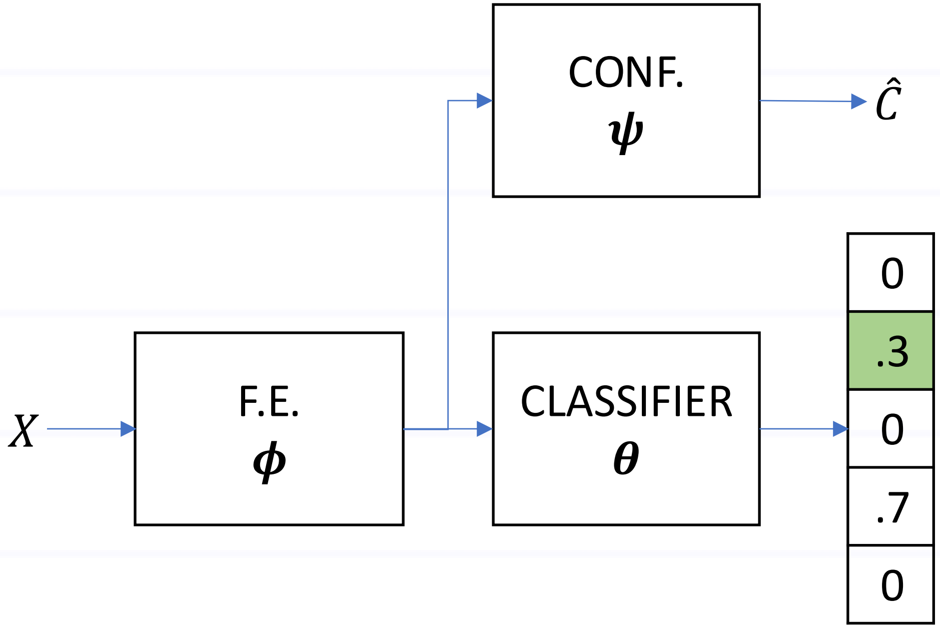
- Introduction
- Why mis-calibration happens?
- Assessing calibration
- Confidence calibration: post hoc methods and Bayesian methods
- Disentangling sources of uncertainty: Epistemic and Aleatoric

How to assess the calibration?

Guo et al., On Calibration of Modern Neural Networks, ICML 2017

Post-hoc calibration

- Calibrate the model <u>after</u> training
- Notations:
- input: X
- ground truth: $Y \in \mathcal{Y} = \{1, \ldots, K\}$
- output: $\hat{Y} = f_\theta\left(f_\phi(X)\right)$
- output class: $\hat{Y}^* = \mathrm{argmax}_k \hat{Y}[k]$
- confidence: $\hat{C} = f_\psi\left(f_\phi(X)\right)$



Calibration

- **Que**: When is the model perfectly calibrated, $P\left(\hat{Y}^* = Y \middle| \hat{C} = c\right) =?$
- $P\left(\hat{Y}^* = Y \middle| \hat{C} = c\right) = c, \forall c \in [0,1]$
- **Que**: what terms above are functions of X?
- $Y, \hat{Y}^*, \hat{C}$
- **Que**: what terms above are independent variables?
- c

Problem

$$P(\hat{Y}^* = Y | \hat{C} = c) = c, \forall c \in [0,1]$$

- c is a continuous variable. How many X in a finite database can have $\hat{C} = c$
- So, we need approximation
- E.g., binning of c

Reliability Diagrams

- Accuracy vs confidence
- Let B_m be the set of samples with $\hat{C}$ falling in mth bin
- Accuracy, $acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i^* = y_i)$
- Confidence, $conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{C}_i$
- For perfectly calibrated model, $acc(B_m) = conf(B_m) \forall m = \{1, \ldots M\}$

Exercise

For a flower classification task,

$$Y = \{RRRR\ JJJJ\ LLLL\}$$

$$\hat{Y} = \{RLRR\ JLJR\ LLRJ\}$$

$$\hat{C} = \{.8, .7, .4, .7, \qquad .7, .8, .8, .4, \qquad .8, .7, .4, .4\}$$

Draw the reliability diagram.

Expected Calibration Error

- Average over the reliability diagram
- ECE $= \mathbb{E}_{\hat{C}}\left[\left|P\left(\hat{Y}^* = Y \middle| \hat{C} = c\right) - c\right|\right]$
- **Que**: write the Monte Carlo approximate of ECE
- ECE $= \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$

Maximum Calibration Error

- Take max error over the reliability diagram
- MCE = $\max\limits_{c\in[0,1]} \left|P\left(\hat{Y}^* = Y \middle| \hat{C} = c\right) - c\right|$
- **Que**: write the Monte Carlo approximate of MCE
- MCE = $\max\limits_{m\in\{1,\ldots,M\}} |acc(B_m) - conf(B_m)|$

Outline

- Introduction
- Why mis-calibration happens?
- Assessing calibration
- Confidence calibration: post hoc methods and Bayesian methods
- Disentangling sources of uncertainty: Epistemic and Aleatoric

Confidence Calibration

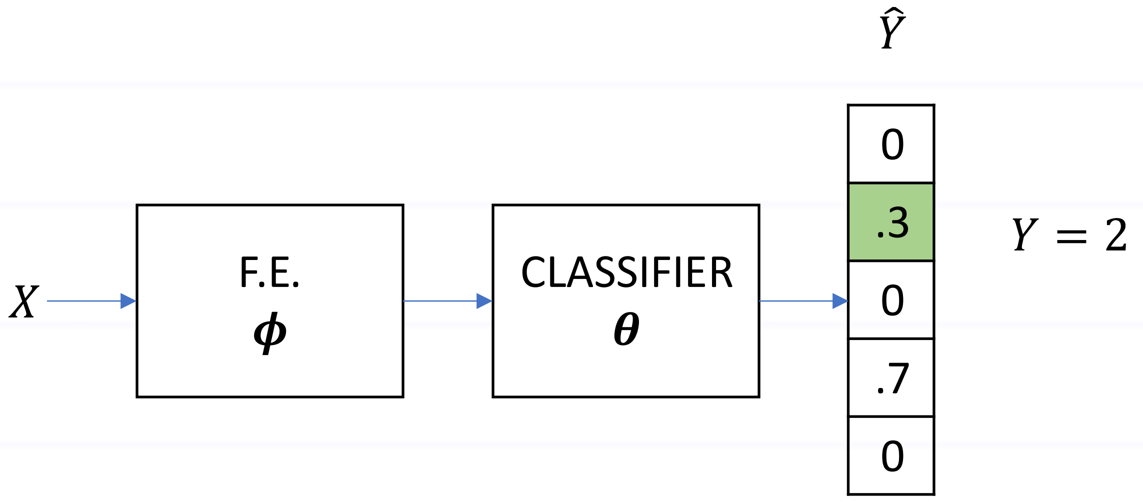
Guo et al., On Calibration of Modern Neural Networks, ICML 2017

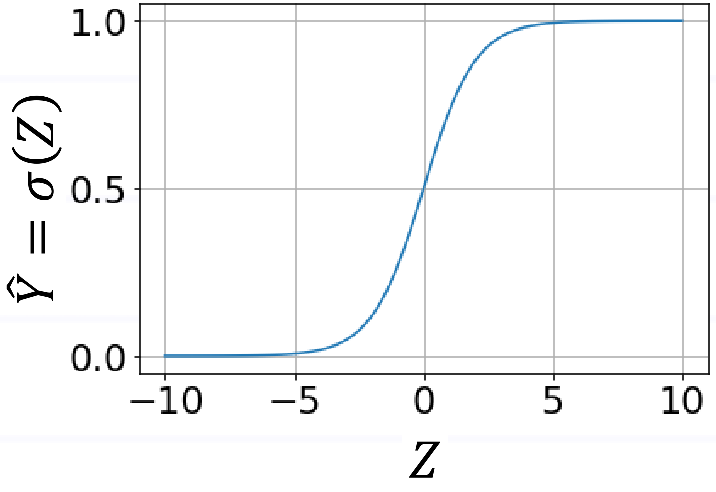
Goal

- Derive $\hat{C}$ using $\hat{Y}, \hat{Y}^*, Z, X$

$$\hat{C}\left(\hat{Y}, \hat{Y}^*, Z, X\right)$$

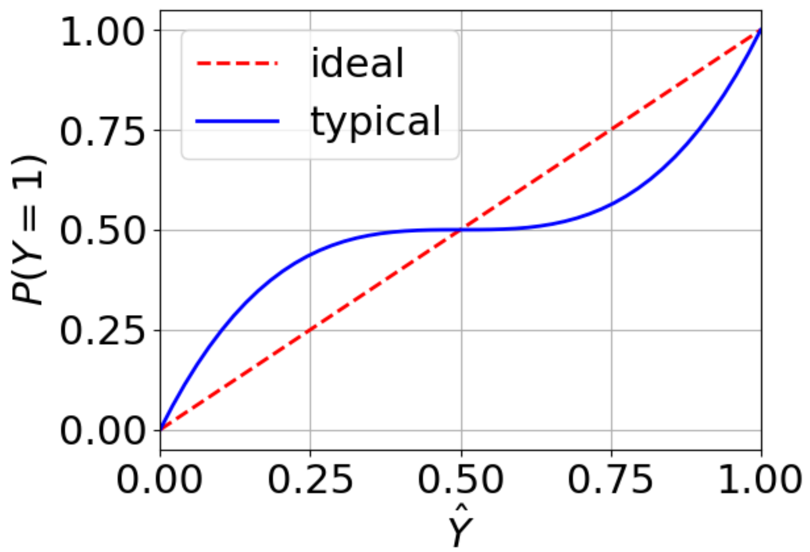
- Can't use Y during testing
- Let's focus on binary classification
- **Que**: Draw $\hat{Y}$ vs Z
- **Que**: Draw typical $\hat{C}\left(\hat{Y}\right)$





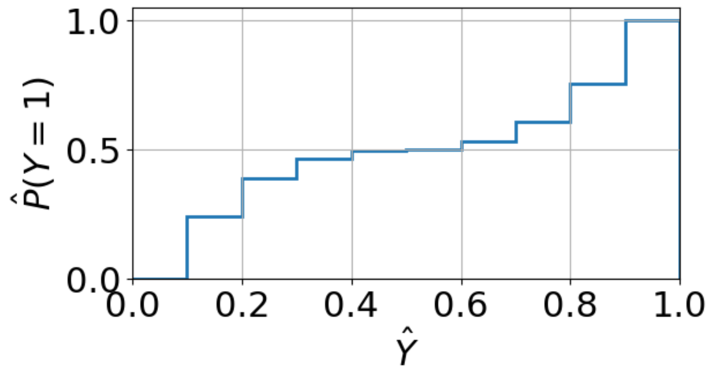
1. Histogram Binning Method

- $\hat{C}(\hat{Y})$
- Instead of modeling $\hat{C}(\hat{Y})$, we can model $\hat{C}(k) := P(Y = k)$ for all k
- **Que**: Draw typical $P(Y = 1)$ vs $\hat{Y}$
- **Que**: What is $\hat{C}(\hat{Y})$ if you know $P(Y = 1)$?
- **Ans**: $\hat{C}(\hat{Y}) = P\left(Y = \arg\max_k \hat{Y}\right)$. Can you draw it for binary case?



MADHAV

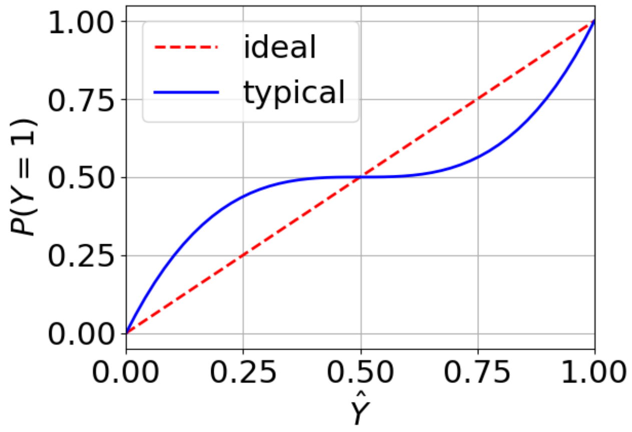
1. Histogram Binning Method



- Divide $\hat{Y} \in [0,1]$ into M bins
- Let $P(Y = 1) = \theta_m$ if $\hat{Y}$ falls in bin m
- To estimate θ_m, Loss $= \sum_m \sum_i 1\left(\hat{Y}_i \in \text{bin } m\right)(\theta_m - Y_i)^2$
- **Que**: What is the optimal θ_m?
- $\theta_m = \frac{\sum_i 1(\hat{Y}_i \in \text{bin } m) Y_i}{\sum_i 1(\hat{Y}_i \in \text{bin } m)}$

2. Isotonic Regression Method

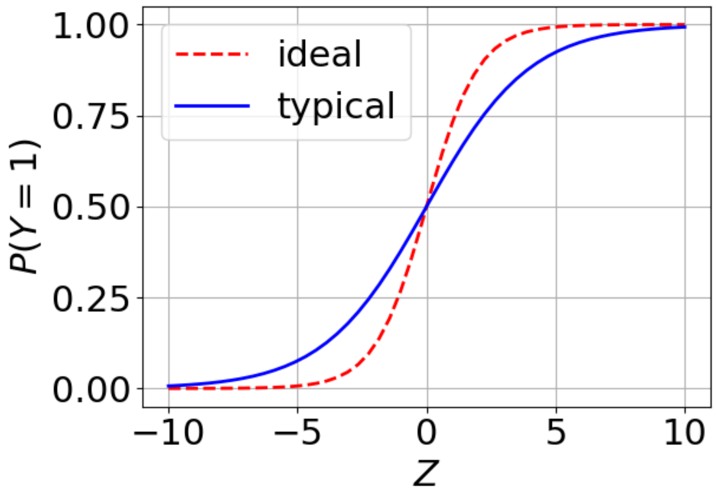
- $\hat{C}(\hat{Y})$
- One can learn $P(Y = 1)$ as a function of $\hat{Y}$ using simple regression models
- Isotonic regression model is one such model



3. Platt Scaling Method

- $\hat{C}(Z)$

- **Que**: Draw typical $P(Y = 1)$ vs Z



- Approximate this with a sigmoid as

$$P(Y = 1) = \sigma(aZ + b)$$

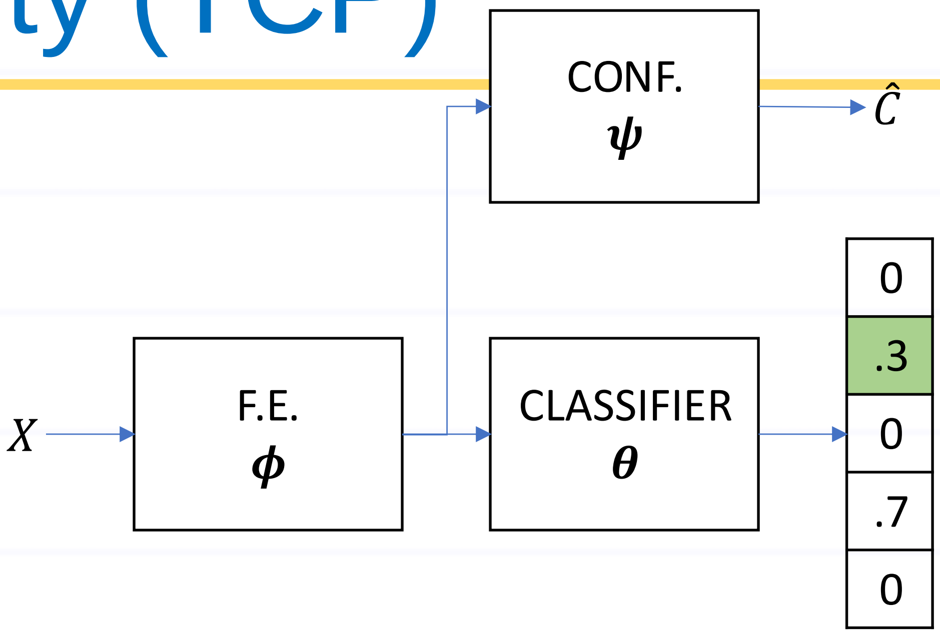
- If $b = 0$, it is called Temperature Scaling method with $a = 1/T$

3. Platt Scaling Method

$$P(Y = 1) = \sigma(aZ + b)$$

- a and b are estimated using MLE over validation data
- **Que**: Could you write the loss function?
- Loss = $-\sum_i \delta_{y_i,1} \ln \sigma(aZ + b) + \delta_{y_i,0} \ln\left(1 - \sigma(aZ + b)\right)$

4. True Class Probability (TCP)



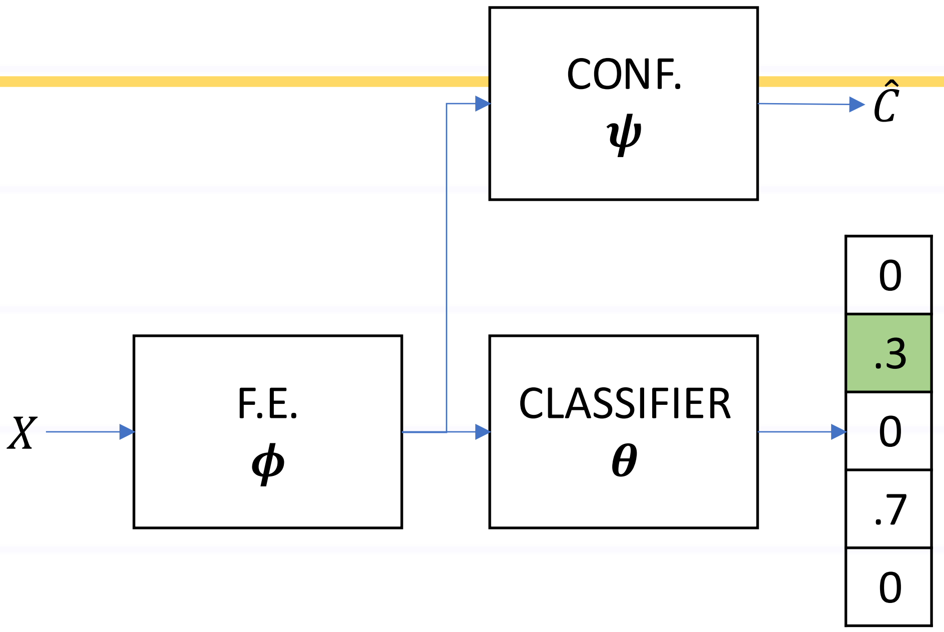
- $\hat{C}(X)$
- Train ψ as a regressor
- Loss $= \mathbb{E}\left[\left(C(X) - \hat{C}(X)\right)^2\right]$
- What should be target?
- Use $C(X) = \hat{Y}[k]$; $k = Y$

Corbiere, et al., "Addressing failure prediction by learning model confidence," NeurIPS 2019.

MADHAV

4. Normalized TCP

- $\hat{C}(X)$
- Loss $= \mathbb{E}\left[\left(C(X) - \hat{C}(X)\right)^2\right]$
- When number of classes is large, $\hat{Y}[k]$ gets smaller
- Use $C(X) = \dfrac{\hat{Y}[k]}{\max_{k'} \hat{Y}[k']}$; $k = Y$

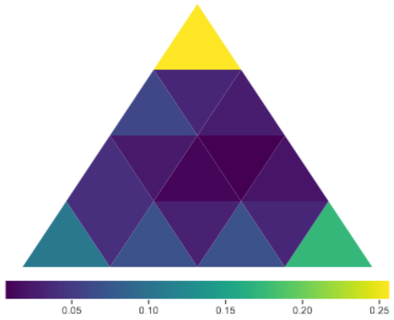


Corbiere, et al., "Addressing failure prediction by learning model confidence," NeurIPS 2019.

MADHAV

For multi-class classification

- Treat it as K one-vs-all problems
- Estimate $P(Y = k)$ for all k using $Z[k]$ or Z
- Hence, get $\hat{C}[k]$
- **Que**: Why not treat it as a full multi-class problem?
- Curse of dimensionality



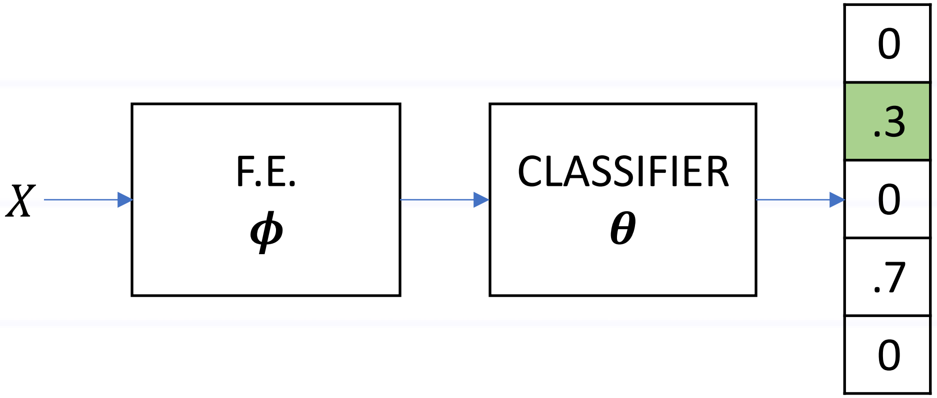
Bayesian Methods

Gal and Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, ICML 2016

Bayesian Neural Network

- Typical NN

$$\hat{Y} = f_\theta\left(f_\phi(X)\right)$$



- In Bayesian NN, ϕ, θ are also random variables; so, we get

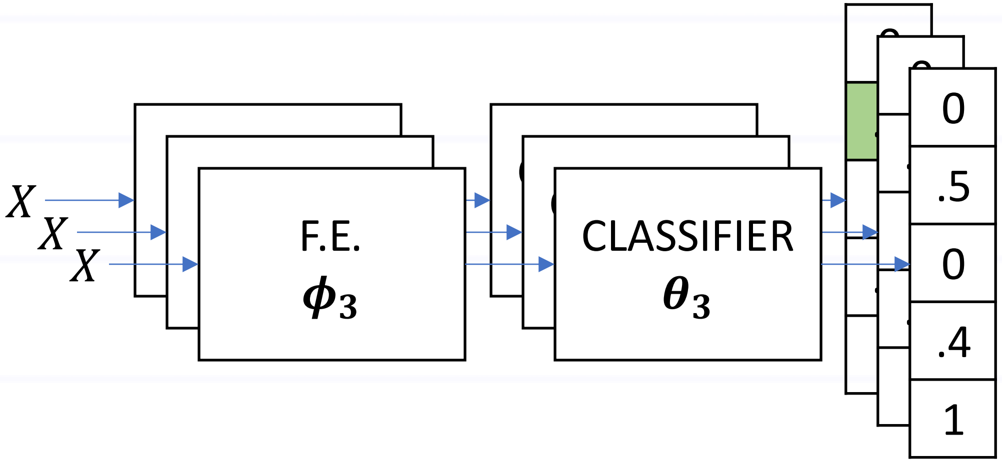
$$p(\hat{Y}|X) = \iint p(\hat{Y}|X, \theta, \phi)p(\theta, \phi)d\theta d\phi$$

Monte Carlo Estimation

- $p(\hat{Y}|X) = \iint p(\hat{Y}|X,\theta,\phi)p(\theta,\phi)d\theta d\phi$
- $p(\hat{Y}|X) = \mathbb{E}_{\theta,\phi\sim p(\theta,\phi)}[p(\hat{Y}|X,\theta,\phi)]$
- $p(\hat{Y}|X) \approx \frac{1}{N}\sum_i p(\hat{Y}|X,\theta_i,\phi_i); \quad \theta_i,\phi_i \sim p(\theta,\phi)$

Monte Carlo Dropouts

- $p(\hat{Y}|X) \approx \frac{1}{N}\sum_{i=1}^{N} p(\hat{Y}|X,\theta_i,\phi_i)$; $\theta_i,\phi_i \sim p(\theta,\phi)$
- $\theta,\phi \sim p(\theta,\phi)$ is approximated using dropout



Uncertainty

- $\mu = \frac{1}{N} \sum_i \hat{Y}_i$
- $\Sigma = \frac{1}{N} \sum_i \left(\hat{Y}_i - \mu\right)^\intercal \left(\hat{Y}_i - \mu\right)$
- Indicators of uncertainty:
  - Total variance $= \sum_{k,l} \Sigma_{kl}$
  - Entropy $= -\sum_c \hat{Y}[c] \ln \hat{Y}[c]$

Ensemble Method

- $p(\hat{Y}|X) \approx \frac{1}{N}\sum_i p(\hat{Y}|X, \theta_i, \phi_i)$; $\theta_i, \phi_i \sim p(\theta, \phi)$
- $\theta, \phi \sim p(\theta, \phi)$ is a model trained using stochastic optimization algorithms
- We train N different models and use them to estimate uncertainty

MADHAV

Input Perturbation Method

- $p(\hat{Y}|X) = \int p(\hat{Y}|\tilde{X})p(\tilde{X}|X)d\tilde{X}$
- $p(\hat{Y}|X) \approx \frac{1}{N}\sum_i p(\hat{Y}|\tilde{X}_i\ )\,;\quad \tilde{X}_i = X + \epsilon_i$
- Here, ϵ_i is noise or systematic perturbation of input X

Outline

- Introduction
- Why mis-calibration happens?
- Assessing calibration
- Confidence calibration: post hoc methods and Bayesian methods
- Disentangling sources of uncertainty: Epistemic and Aleatoric

Sources of Uncertainty

- Kendall and Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, NeurIPS 2017

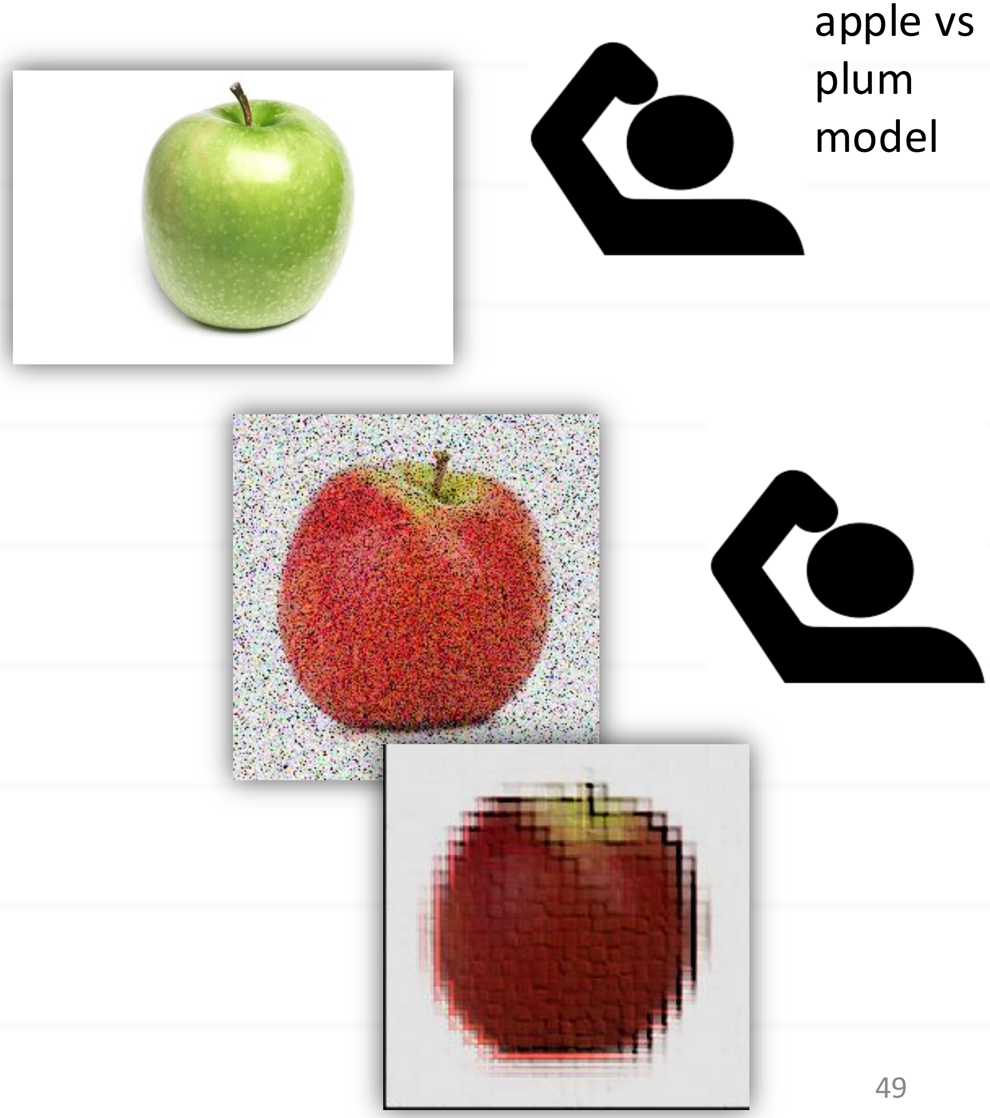
- Hüllermeier and Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Machine Learning 2021

Calibration is not enough

Two kinds of uncertainty:

1. **Model** is limited, not trained on this data or class. **Epistemic Uncertainty**

2. **Data** ambiguous, even if model has been trained on similar data. **Aleatoric Uncertainty**



Should we distinguish?

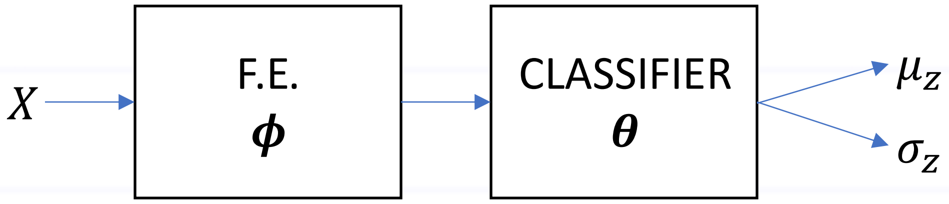
- Epistemic Uncertainty
  - tells about out of domain data (new, unseen inputs) (useful for model adaptation and active learning)
  - tells about anomalies and outliers
  - tells about unseen classes (novel class discovery)
- Aleatoric Uncertainty
  - tells about difficult data which needs manual intervention. More training won't help.

Can we distinguish?

Yes, two approaches

1. Bayesian NN
2. Evidential learning [Sensoy et al., Evidential Deep Learning to Quantify Classification Uncertainty, NeurIPS 2018]

Consider the Z space (logits)



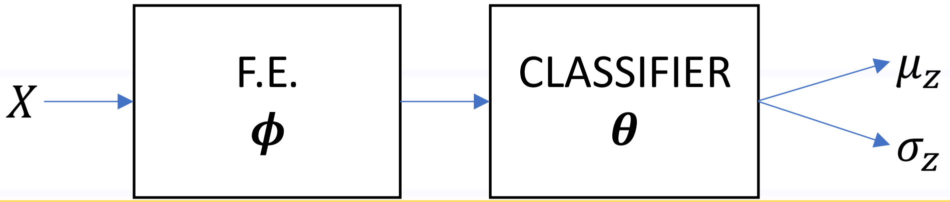
- Assume Gaussian output

$$Z \sim \mathcal{N}(\mu_z, \sigma_z^2); \mu_z \in \mathbb{R}^K, \sigma_z \in \mathbb{R}^K$$

where $\mu_z, \sigma_z = f_\theta\left(f_\phi(X)\right)$

- Output is as usual $\hat{Y} = \text{softmax}(Z)$ for classification and $\hat{Y} = Z$ for regression

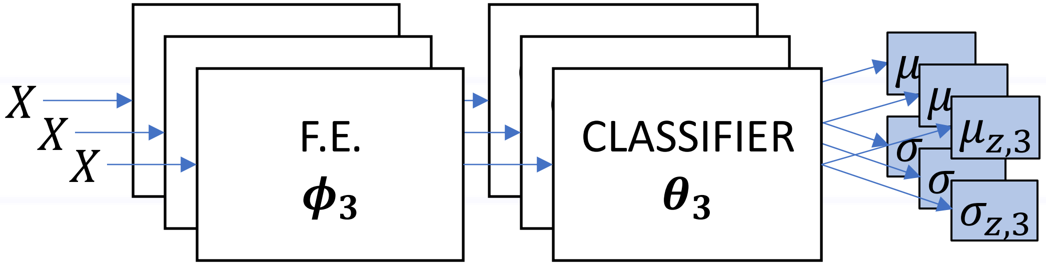
Uncertainty



- σ_Z quantifies the uncertainty in Z, but what kind of uncertainty?
- Let us perturb the model parameters
- $\mu_{Z,i}, \sigma_{Z,i} = f_{\theta_i}\left(f_{\phi_i}(X)\right)$

Uncertainty

- Que: What is $\mathbb{E}[Z]$?



$$\mathbb{E}[Z] = \iint Zp(Z|X,w)p(w)dw\, dZ \;\; ; \quad w = \{\phi, \theta\}$$

$$= \int \mu_{z,w}\; p(w)dw$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \mu_{z,i}$$

Uncertainty

- Que: What is $\mathrm{covar}[Z]$ or $\mathbb{E}[(Z - \mathbb{E}[Z])^\intercal (Z - \mathbb{E}[Z])]$?

$$\mathbb{E}[Z^\intercal Z] = \iint Z^2 p(Z|X,w)p(w)dw\, dZ$$

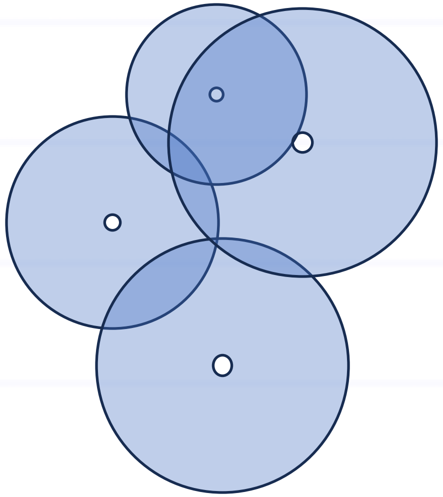
$$= \int \left(\sigma_{z,w}^2 I + \mu_{z,w}^\intercal \mu_{z,w}\right) p(w)dw$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_{z,i}^2 I + \mu_{z,i}^\intercal \mu_{z,i}\right)$$

Uncertainty

- $\mathrm{covar}[Z] \approx \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_{z,i}^2 I + \mu_{z,i}^\intercal \mu_{z,i}\right) - \mathbb{E}[Z]^2$
- $= \underbrace{\frac{1}{N}\sum_i \sigma_{z,i}^2 I}_{\textbf{Aleatoric Uncertainty}} + \underbrace{\frac{1}{N}\sum_i \mu_{z,i}^\intercal \mu_{z,i} - \left(\frac{1}{N}\sum_{i=1}^{N} \mu_{z,i}\right)^2}_{\textbf{Epistemic Uncertainty}}$

PICTORIAL UNDERSTANDING



How to Train? (for regression)

- $\mathcal{L}(w) = -\mathbb{E}[\ln p(Z|X,w)]$ max likelihood
- **Que**: What is it with Gaussian $p(Z|X,w)$
- $\mathcal{L}(w) = \mathbb{E}\left[\sum_k \left(\frac{1}{2}\ln 2\pi\sigma_k^2 + \frac{(\mu_k - Z_k)^2}{2\sigma_k^2}\right)\right]$
- Here, μ_k, σ_k are NN outputs and Z_k is the ground truth; k is dim
- $\mathcal{L}(w) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_k \left(\frac{1}{2}\ln 2\pi\sigma_{k,i}^2 + \frac{(\mu_{k,i} - Z_k)^2}{2\sigma_{k,i}^2}\right)\right]$ with perturbations of w; the $\mathbb{E}$ is over samples of X

How to Train? (for classification)

- $\mathcal{L}(w) = -\mathbb{E}\left[\ln p\left(\hat{Y}^* \middle| X, w\right)\right]$ max likelihood or X-entropy
- Let us perturb Z to obtain $\hat{Y}$
- $Z_t = \mu_{z,w} + \sigma_{z,w}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I), t = 1, \dots, T$
- $\mathcal{L}(w) = -\mathbb{E}\left[\ln\left(\frac{1}{T}\sum_t p\left(\hat{Y}^* \middle| X, w, Z_t\right)\right)\right]$
- **Que**: What is it with softmax?
- $\mathcal{L}(w) = -\mathbb{E}\left[\ln\left(\frac{1}{T}\sum_t \frac{e^{Z_{t,k^*}}}{\sum_{k'} e^{Z_{t,k'}}}\right)\right]$
- Here, k^* is the ground truth class; the $\mathbb{E}$ is over samples of X and perturbations of w

[Kendall & Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?]

Further Reading

- Lakshminarayanan et al., "*Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*," NeurIPS, 2017
- Seitzer, et al., "*On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks*," ICLR 2022
- Sensoy et al., "*Evidential Deep Learning to Quantify Classification Uncertainty*", NeurIPS 2018
- Ryan Tibshirani, "*Conformal Prediction*", Advanced Topics in Statistical Learning, Spring 2023

Questions?

Next: Part 2