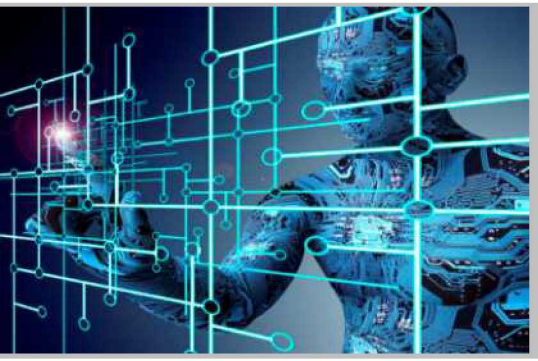
SAND2020-1792PE



Designing for Interpretability and Adaptability by Using Weighted Averages

Sapan Agarwal, Justin Wong, Andrew De La Cruz

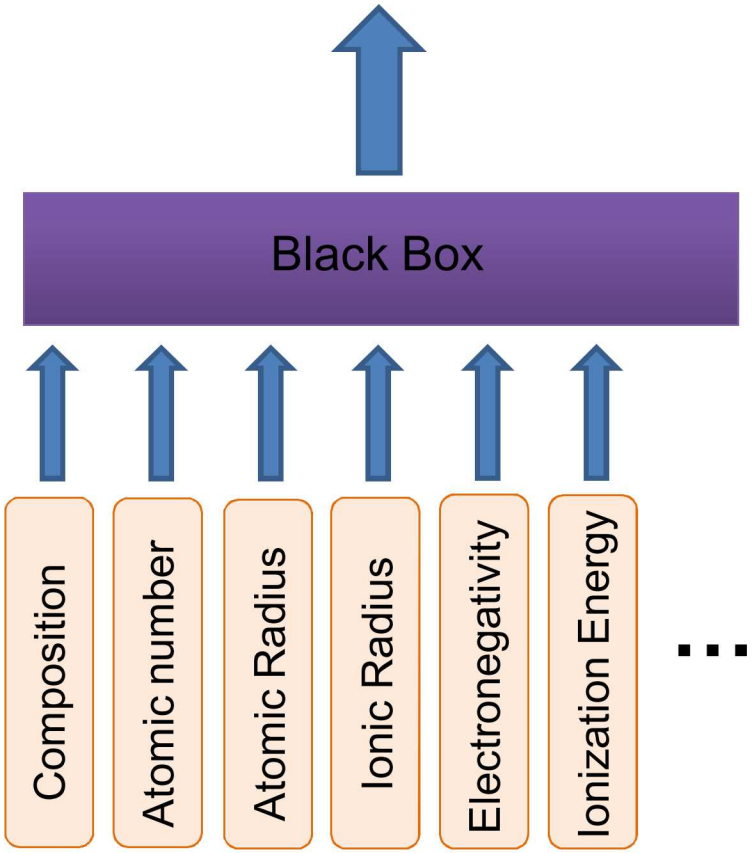
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

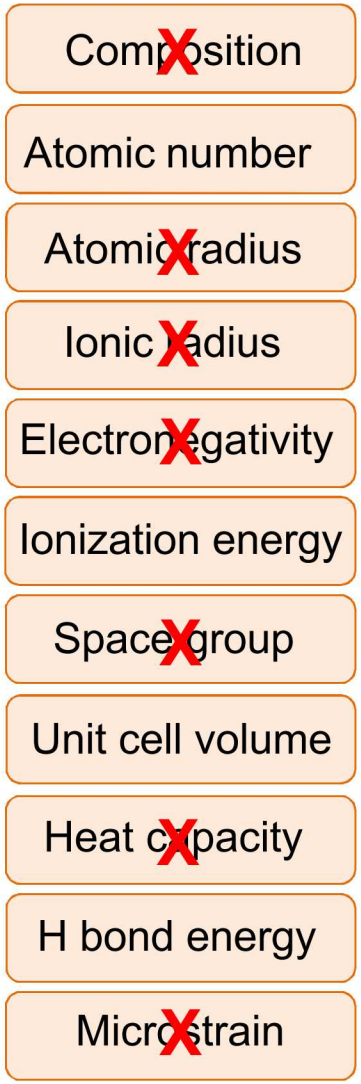
Conventional vs explainable machine learning

Sandia National Laboratories

Conventional ML

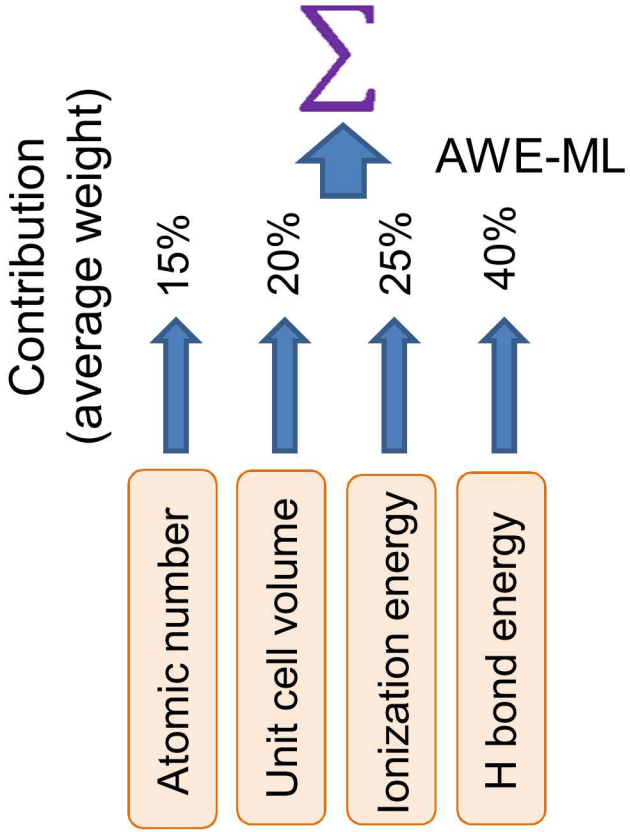
Observable: *Metal reacts with H_2 at SPT*





Explainable ML

Observable: *Metal reacts with H_2 at SPT*



⇒ ***The EML framework has the potential to address exceedingly complex interrelationships among features that defy scientific intuition to extract structure-property relationships***

Explainability background

- Explainability and accuracy are frequently at odds in classification and ranking problems
- Logistic regression is imminently explainable
- $\boldsymbol{logit(C) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i}$
- However…for even trivially simple examples, like MNIST (handwriting), LR scores about 93% accuracy vs record accuracies around 99.8%

Ideal classifier

1. Highly accurate
2. Explains each instance
3. Streaming, modifiable on the fly
4. Scalable

Use our new classifier:
AWE-ML: Averaged Weights for Explainable Machine Learning

Directly Use Information from Training Data

Sandia National Laboratories

Find a value Z given measured features

$$x_1 = x_1', x_2 = x_2', x_3 = x_3'$$

i.e. estimate $Z|\, x_1 = x_1', x_2 = x_2', x_3 = x_3'$

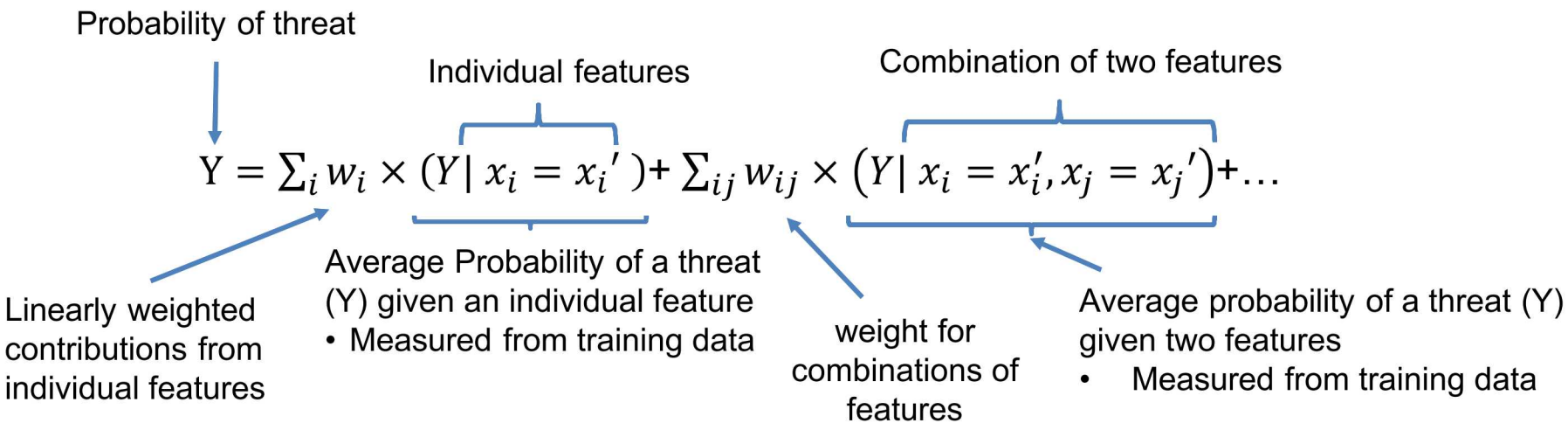
From training data, measure the following values:

$$< Z >= 0.8$$
$$< Z|\, x_1 = x_1' \geq -0.001$$
$$< Z|\, x_2 = x_2' > = 0.9$$
$$< Z|\, x_3 = x_3' > = 0.7$$
$$<Z|\, x_1 = x_1', x_2 = x_2' > = 2$$

How do we combine these to get the best overall estimate?

"Averaged Weights for Explainable Machine Learning" (AWE-ML) algorithm

AWE-ML predicts a value Y (e.g. probability of a threat) given features x_i as follows:



“Averaged Weights for Explainable Machine Learning” (AWE-ML) algorithm

- **Algorithm was initially derived as a set of heuristic rules**
  - **Working towards mathematically formalizing the model.**
- **Empirically, heuristic rules have accuracies matching state of the art**

Heuristic Model	Bayesian Model	Measure Theory Model
Write a series of heuristic rules to estimate value, accounting for noise, combinations of features, and standard deviation of values measured from the training data	Treat each feature or combination of features as evidence that updates the prediction	Use measure theory to justify and improve heuristic model equations

Heuristic Model to Estimate Probability Y =P(Z=1)

- One route to combine features is by averaging the probabilities

$$P(Z=1|\, x_1 = x_1{}', x_2 = x_2{}', x_3 = x_3{}') \approx \frac{\sum_i P(Z=1|\, x_i = x_i{}')}{\sum_i 1}$$

- Need to account for many effects:
  - Use combinations of features: $P\left(Z=1|\, x_i = x_i', x_j = x_j'\right)$
  - As a P approaches 0 or 1, it should be given more weight
  - Measured probabilities with less supporting data should be given less weight
  - Measured probabilities that give new information should possibly be given more weight
  - Account for class imbalance

$$P = \sum_i w_i \times P(Z=1|\, x_i = x_i{}') + \sum_{ij} w_{ij} \times P\left(Z=1|\, x_i = x_i', x_j = x_j{}'\right) + \ldots$$

Use a Hierarchical Weighted Average

Bayesian Model

- We want to estimate Y given measured features $x_1 = x_1', x_2 = x_2', x_3 = x_3', \ldots$
- Model possible values of Y as a normal distribution with mean μ_y that we are trying to estimate.
- Treat individual feature values in a particular instance as data points that we are using to update the estimate of μ_y.
  - Data point$_i = Y = \mathrm{Y} | x_i = x_i'$
- Bayes Rule:

posterior likelihood prior

$$P\left(\mu_y \,\middle|\, \cap_i \; Y = Y_i\right) \propto P(\cap_i \; Y = Y_i | \mu_Y) \; P(\mu_Y)$$

Limitations: Assumes normal distributions

Accuracy

Dataset	AWE-ML	Linear SVM	SVM (RBF)	Logistic Regression	Random Forests
Bank	$90.1\% \pm 0.3\%$	$90.1\% \pm 0.4\%$	$89.2\% \pm 0.1\%$	$90.1\% \pm 0.3\%$	$90.4\% \pm 0.2\%$
Breast Cancer	$96.7\% \pm 1.9\%$	$96.5\% \pm 1.4\%$	$96.9\% \pm 2.5\%$	$96.7\% \pm 2.3\%$	$95.1\% \pm 1.9\%$
Credit	$82.2\% \pm 0.3\%$	$79.7\% \pm 0.6\%$	$82.0\% \pm 0.6\%$	$81.1\% \pm 0.3\%$	$81.6\% \pm 0.4\%$
Customer	$92.0\% \pm 2.3\%$	$91.1\% \pm 2.3\%$	$92.5\% \pm 3.9\%$	$90.2\% \pm 4.9\%$	$92.2\% \pm 2.5\%$
Iris	$98.0\% \pm 3.0\%$	$96.0\% \pm 6.1\%$	$96.0\% \pm 4.4\%$	$95.3\% \pm 5.2\%$	$96.0\% \pm 4.4\%$
Lymphography	$85.5\% \pm 11.9\%$	$85.1\% \pm 12.9\%$	$83.7\% \pm 10.5\%$	$83.7\% \pm 7.8\%$	$84.4\% \pm 3.9\%$
Promoter	$94.3\% \pm 6.4\%$	$93.3\% \pm 6.0\%$	$92.4\% \pm 5.7\%$	$92.4\% \pm 7.0\%$	$91.5\% \pm 10.8\%$
Spect	$83.0\% \pm 6.0\%$	$82.0\% \pm 4.9\%$	$79.4\% \pm 1.3\%$	$79.4\% \pm 1.3\%$	$81.2\% \pm 5.7\%$
Splice	$96.4\% \pm 0.9\%$	$95.0\% \pm 1.2\%$	$86.4\% \pm 1.1\%$	$95.9\% \pm 0.9\%$	$94.6\% \pm 1.1\%$
Transfusion	$78.0\% \pm 2.7\%$	$77.1\% \pm 2.0\%$	$76.0\% \pm 0.3\%$	$76.7\% \pm 2.4\%$	$73.1\% \pm 2.7\%$
Voting Records	$96.7\% \pm 1.8\%$	$94.4\% \pm 3.5\%$	$95.1\% \pm 2.1\%$	$95.4\% \pm 3.0\%$	$95.1\% \pm 2.1\%$
Average	90.3%	89.1%	88.1%	88.8%	88.65%

Dataset	AWE-ML	Random Forests
Biofuels - Cetane	89.4% ± 3.5%	89.8%± 2.9%
Biofuels – Octane Sensitivity	84.0% ± 6.4%	80.2 ± 5.2%

Hyper parameters for all classifiers optimized using 4 fold cross validation

Use Model to Analyze a Misclassified Result

Sandia National Laboratories

1984 Congressional Voting Records Dataset

The classifier predicted with 70% probability that this Member of Congress would be a Republican when they are a Democrat.

Features	Counts Rep.	Counts Dem.	Probability Rep.	Weight	Cumulative Weight
Immigration-Y, South Africa Export Act-N	16	0	100%	19.4%	19.4%
Doc Fee-Y, Mx Missile-N, Immigration-Y, Duty Free-N	47	0	100%	4.8%	24.2%
Doc Fee-Y, Contras aid-N, Immigration-Y, Duty Free-N	47	0	100%	4.7%	28.9%
Adopt Budget-Y, Synfuels cutback-Y, Education-N	0	58	0%	4.0%	32.9%
Water-Y, Adopt Budget-Y, Synfuels cutback-Y,	0	38	0%	3.1%	36.0%

Feature #	Probability Republican	Weight	Feature #	Probability Republican	Weight
Immigration-Y	95%	17%	Mx missile-N	86%	4%
Doc fee freeze-Y	94%	16%	Nicaraguan contra aid – N	83%	4%
South Africa Export Admin Act-N	94%	14%	Anti-satellite test ban-N	29%	3%
Adopt Budget-Y	14%	8%	Handicapped infants-N	78%	3%
Synfuels corporation cutback-Y	23%	8%	Crime-Y	55%	2%
Education spending-N	21%	6%	El Salvador aid-Y	69%	2%
Duty free exports-N	89%	5%	Superfund right to sue-Y	52%	2%
Water project cost sharing-Y	53%	4%	Religious groups in schools-Y	56%	2%

Republican features have higher certainty and therefore higher weight

Probability given all feature combinations containing the specified feature

Analyze a Correctly Classified Result

Sandia National Laboratories

1984 Congressional Voting Records Dataset

The classifier correctly predicted this member of congress is a republican

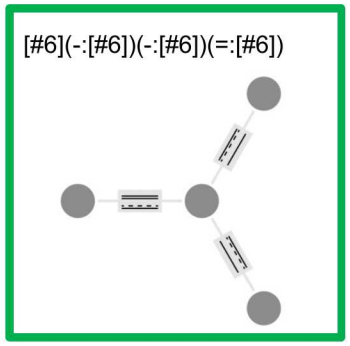
Features	Counts Rep.	Counts Dem.	Probability Rep.	Weight	Cumulative Weight
Budget -N, Doc Fee-Y, Immigration-Y	54	0	100%	18.3%	18.3%
Doc Fee-Y, Immigration-Y, Education –Y	57	0	100%	16.1%	34.4%
Immigration-Y, Education-Y, SA Export-N	16	0	100%	5.0%	39.4%
Water-N, Doc Fee-Y, Immigration-Y	31	0	100%	3.0%	42.4%
Water-N, Immigration-Y, SA Export-N	6	0	100%	0.9%	43.3%
Handicap-N, Budget-N, Doc Fee-Y, MxMissile-N, Superfund-Y	76	1	98.7%	0.5%	43.8%

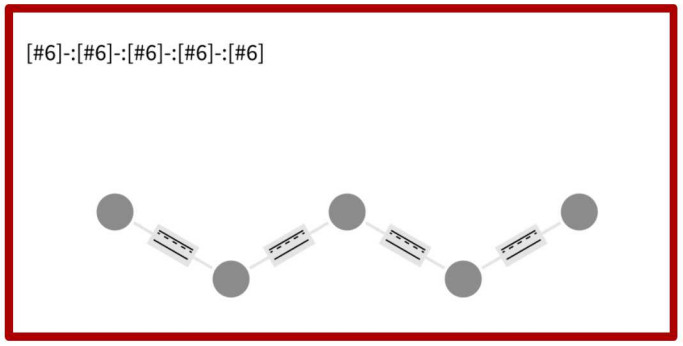
Predicting Biofuels with High Octane Sensitivity

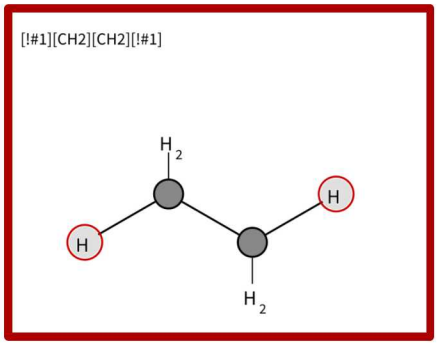
What molecular structures in biofuels result in high octane sensitivity and high Research Octane Number (RON)?

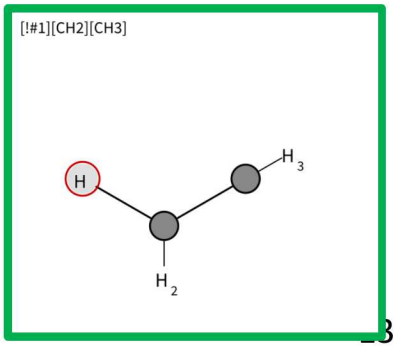
Feature 1	Feature 2	Feature 3	Feature 4	Mean Weight	# Low	# High
[#6](-:[#6])(-:[#6])(=:[#6])	[#6]-:[#6]-:[#6]-:[#6]-:[#6]	[!#1][CH2][CH2][!#1]	[!#1][CH2][CH3]	62%	0	5
[#6]>5				44%	0	7
[#6][#6]1[#6]([#6])[#6][#6][#6]1	[!#1][CH]([!#1])[CH]([!#1])[!#1]			44%	0	4
[#6][#6]1[#6]([#6])[#6][#6][#6]1	[!#1][CH2][CH]([!#1])[CH3]			44%	0	4
[!#1]C(=O)[CH3]				41%	9	0
[#6]-:[#6]-:[#6]-:[#6]-:[#6]-:[#6](-:[#6])-:[#6]	[!#1]c1[cH][cH][cH]c([!#1])c1[!#1]	[!#1][CH2][CH2][!#1]		40%	18	0
CCCCC	CCCCCC	[CD2H](=*)-*		36%	15	0
[#6]-:[#6]-:[#6]-:[#6]-:[#6]-:[#6](-:[#6])-:[#6]	[!#1]c1[cH][cH][cH]c([!#1])c1[!#1]	CCCCCC		34%	17	0

Red = not present
Green = present

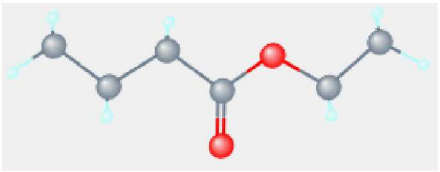






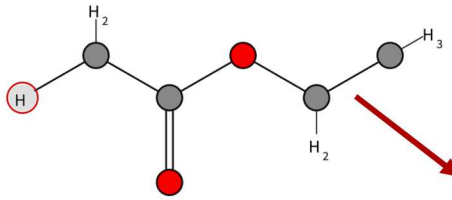


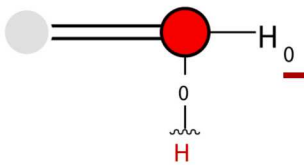
Analyze a particular molecule: Ethyl butyrate $C_6H_{12}O_2$

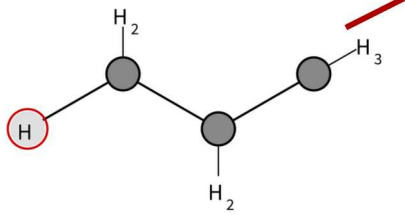


Correctly predicted to have high sensitivity

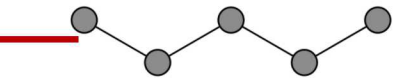
F0	F1	F2	Count Low Sensitivity	Count High Sensitivity	Weight
[OD1H0]=*	[#6]-:[#6]-:[#6]-:[#6]-:[#6]	[!#1][CH2]C(=O)O[CH2][CH3]	0	2	54%
[!#1][CH2]C(=O)O[CH2][CH3]	CCCCC		0	2	18%
[#8]-:[#6]-:[#6]-:[#6]-:[#6]-:[#6]	[!#1][CH2]C(=O)O[CH2][CH3]		0	2	2%







Feature	Weighted Probability	Original Probability	Weight
[!#1][CH2]C(=O)O[CH2][CH3]	100%	67%	32%
[#6]-:[#6]-:[#6]-:[#6]-:[#6]	100%	58%	20%
[OD1H0]=*	100%	19%	19%
CCCCC	100%	51%	9%
[!#1][CH2][CH2][CH3]	100%	10%	3%



Probability given other features present in the molecule (i.e. context dependent)

Probability given this feature only (independent of the context)

Summary

- Matches state of the art machine learning accuracy
  - Identical on open datasets
- Fully explainable
- Coded in Cython for speed
- Compatible with Sci-kit Learn
- To do:
  - Finish developing mathematically rigorous model
  - Parallelize code
  - Adapt to new data without retraining

$$Y = \sum_i w_i \times (Y|\, x_i = x_i{}' \,) + \sum_{ij} w_{ij} \times \left(Y|\, x_i = x_i', x_j = x_j{}'\right) + \ldots$$

Backup

Sandia National Laboratories

Machine Learning 101

- Statistical Model

$$\Pr(X, Y)$$

- Goal

$$\text{Find } f: X \to Y$$

- How?

$$\min_f \mathrm{E} \underbrace{L\big(Y, f(X)\big)}_{\text{loss function}}$$

Ex: L_2 Loss

- Loss Function

$$L\big(Y, f(X)\big) = [Y - f(X)]^2$$

- Minimize expected loss!

$$\min_f \mathrm{E}[Y - f(X)]^2$$

$$\Rightarrow \boxed{f(x) = \mathrm{E}_{Y|X}(Y|X = x)}$$

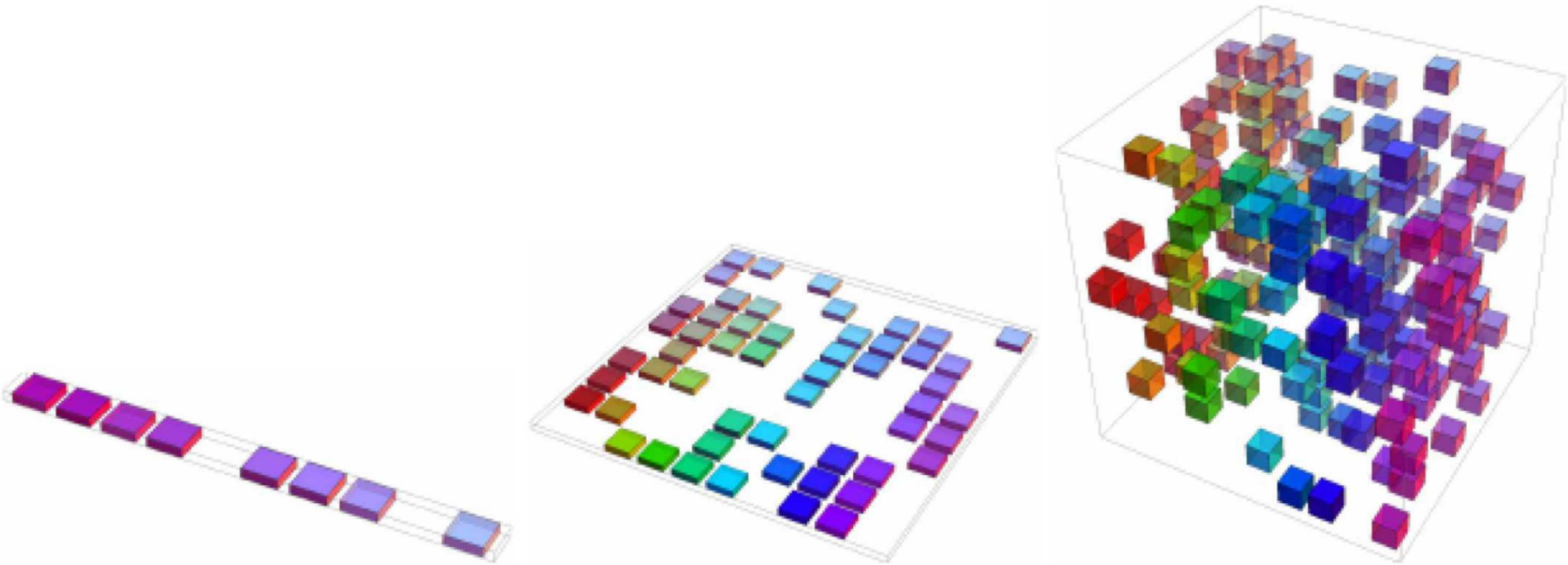
"Regression Function"

Curse of Dimensionality

$$f(x) = \mathrm{E}_{Y|X}(Y|X = x) \approx \mathrm{Ave}(y_i|x_i \text{ near } x)$$

??? (near)

Samples (y_i, x_i) Input (x)



I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning", *The MIT Press* (2016).

Averaged Weights for Explainable Machine Learning (AWE-ML)

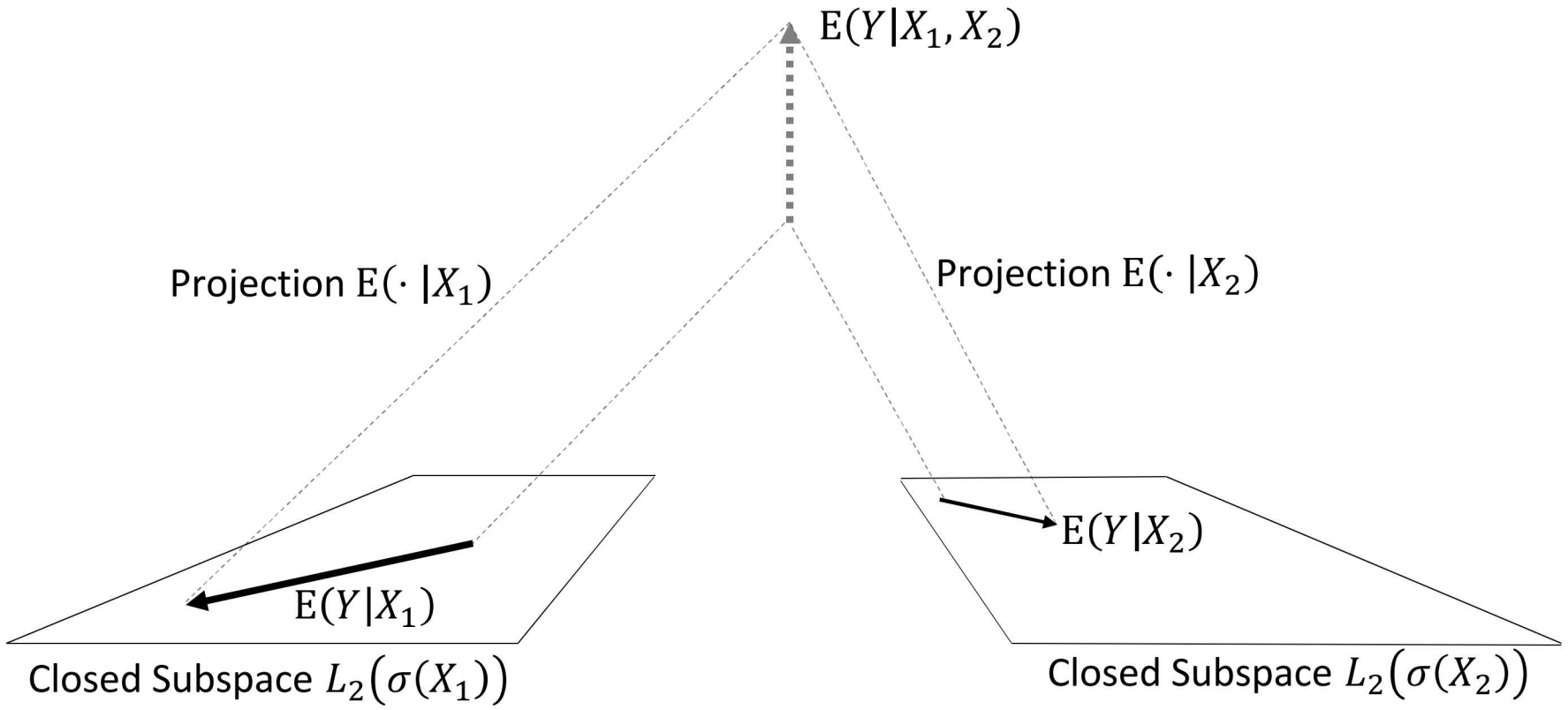
- Assume that we can write

$$\mathrm{E}_{Y|X}(Y|X = x_0) = \sum_j w_j \underbrace{\mathrm{Ave}\left(y_i \middle| x_{ij} = x_{0j}\right)}_{\text{individual features}}$$

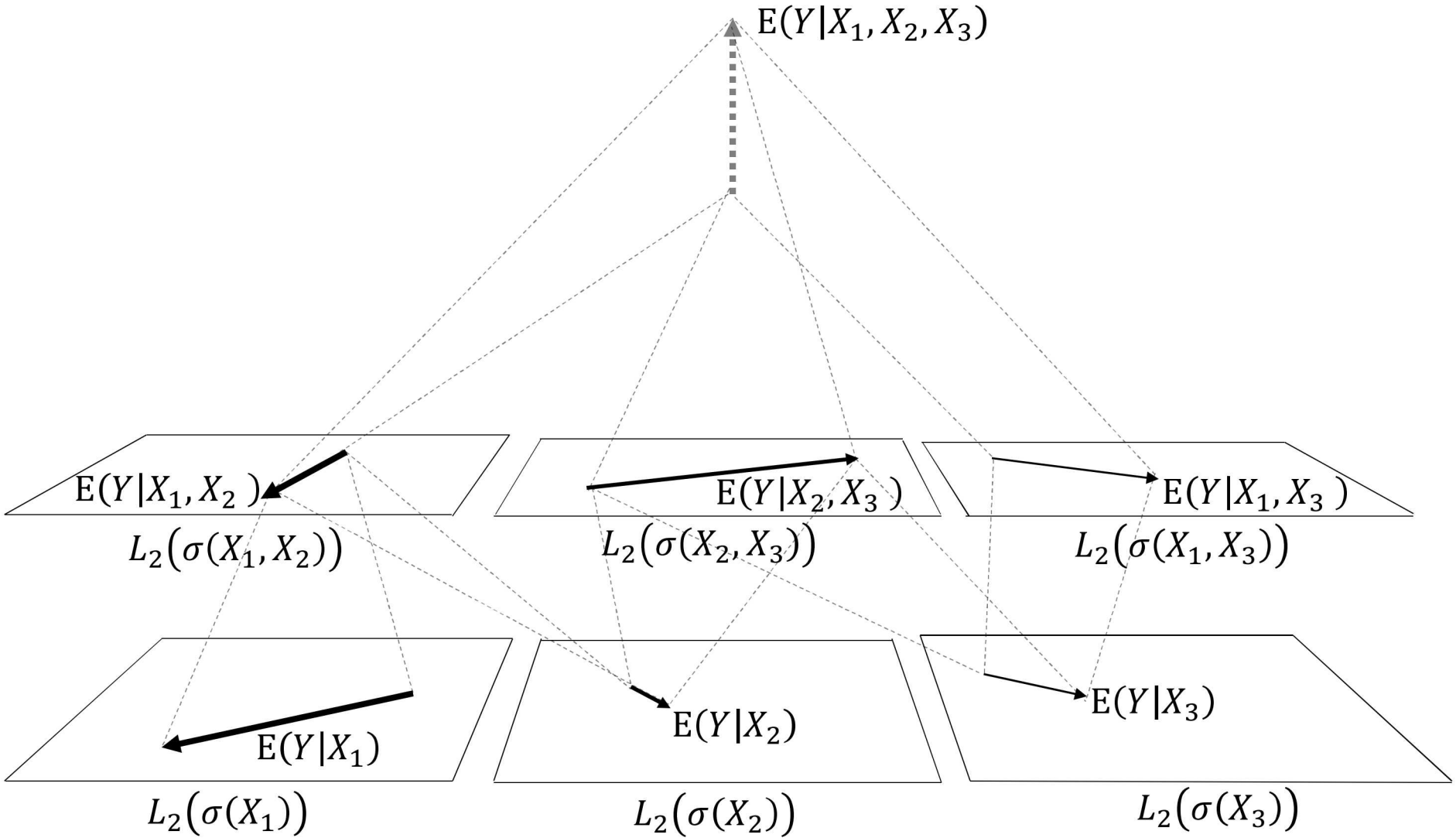
$$+ \sum_{j,k} w_{jk} \underbrace{\mathrm{Ave}\left(y_i \middle| x_{ij} = x_{0j}, x_{ik} = x_{0k}\right)}_{\text{pairs of features}}$$

$$+ \cdots$$

Functional Analysis



Functional Analysis



Averaged Weights for Explainable Machine Learning (AWE-ML)

- Assume that we can write

$$\mathrm{E}_{Y|X}(Y|X = x_0) = \sum_j w_j \underbrace{\mathrm{Ave}\left(y_i \middle| x_{ij} = x_{0j}\right)}_{\text{individual features}}$$

$$+ \sum_{j,k} w_{jk} \underbrace{\mathrm{Ave}\left(y_i \middle| x_{ij} = x_{0j}, x_{ik} = x_{0k}\right)}_{\text{pairs of features}}$$

$$+ \cdots$$

Use Bayesian Updating to Create Weighted Averages for Regression

- We want to estimate Z given measured features $x_1 = x_1', x_2 = x_2', x_3 = x_3', \ldots$
- Model possible values of Z as a normal distribution with mean μ_z that we are trying to estimate
- Treat individual feature values in a particular instance as data points that we are using to update the estimate of μ_z.
  - Data point$_i$ = $Z_i = Z|x_i = x_i'$
- Bayes Rule:

posterior likelihood prior

$$P(\mu_z \mid \cap_i \; Z = Z_i) \propto P(\cap_i \; Z = Z_i | \mu_z) \; P(\mu_z)$$

Using Categorical / Binned Features

Sandia National Laboratories

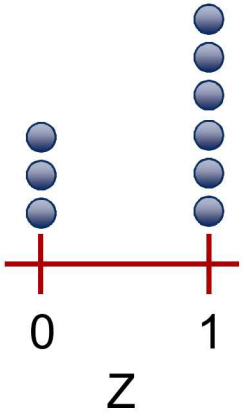
- AWE-ML requires data to be binned into categorical variables
- Binned data allows classifications to be directly tied to specific bins in the training data
- Random forests directly use continuous data averaged over different cuts of the data, preventing identification of the specific subset of the training data that is relevant to classification.
- Binning may result in a small loss in accuracy, but a gives a large improvement in explainability and adaptability.

Likelihood

$$P(\cap_i\ Z = Z_i|\mu_z) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(Z_i-\mu_z)^2}{2\sigma_i^2}}$$

$Z_i = <Z|x_i = x_i'>$

Training data with feature i = i'



i' = feature value in a particular example being classified

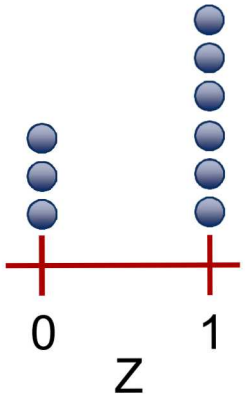
Z_i =mean of data

σ_i^2 = variance of data

Evidence is mean value of Z_i, not the individual data points

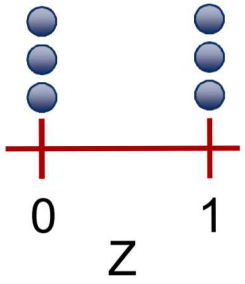
Prior

Feature i=Bin i_1



Z_i^1 =mean of training data with x_i=1

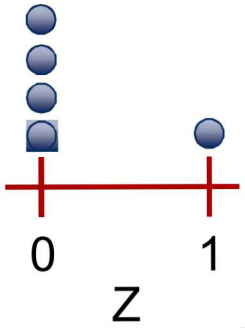
Feature j=Bin j_1



Z_j^1 =mean of training data with x_j=1

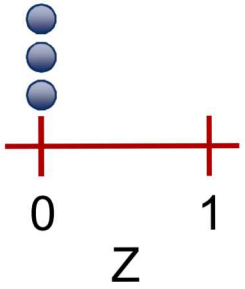
. . .

Feature i=Bin i_2



Z_i^2 =mean of training data with x_i=2

Feature j=Bin j_2



Z_j^2 =mean of training data with x_j=2

$$P(\mu_Z) = \frac{1}{\sqrt{2\pi\tau^2}} e^{\frac{-(\mu_Z - Z^0)^2}{2\tau^2}}$$

Z^0 = average Z over all training data

τ = weighted standard deviation of all possible Z_i in training data, with each feature equally weighted

Derive a Weighted Average from Bayes Rule

posterior likelihood prior

- $P(\mu_z \mid \cap_i\ Z = Z_i) \propto P(\cap_i\ Z = Z_i \mid \mu_z) \times P(\mu_z)$

Assume features are independent

- $P(\mu_z \mid \cap_i\ Z = Z_i) \propto \prod_i P(Z = Z_i \mid \mu_z) \times P(\mu_z)$

Assume normal distributions

Likelihood

$$\prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(Z_i-\mu_z)^2}{2\sigma_i^2}}$$

Prior

$$\frac{1}{\sqrt{2\pi\tau^2}} e^{\frac{-(\mu_z - Z^0)^2}{2\tau^2}}$$

$Z_i =< Z \mid x_i = x_i' >$

σ_i =standard deviation (SD) of data that led to z_i

Z^0 = average Z over all training data

τ = SD of all possible Z_i in training data

Use Bayes Rule to Find Posterior Mean as a Weighted Average

posterior likelihood prior

$$P(\mu_z \mid \cap_i\ Z = Z_i) \propto P(\cap_i\ Z = Z_i \mid \mu_z)\, P(\mu_z)$$

$$\mu_z = \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_n^2}} \times \left(\frac{1}{\tau^2} M_Z + \frac{1}{\sigma_1^2} Z_1 + \frac{1}{\sigma_2^2} Z_2 + \ldots + \frac{1}{\sigma_n^2} Z_n \right)$$

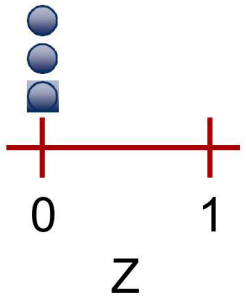
$$\mathrm{Z_i} = <Z|x_i = x_i' >$$

$$\sigma_i = \sigma(Z|x_i = x_i')$$

$$\sigma_z = \frac{\tau^2 \prod_i \sigma_i^2}{\prod_i \sigma_i^2 + \tau^2 \times (\sum_j \prod_{i \neq j} \sigma_i^2)}$$

Compensate for Noise/Low Data counts

Feature i=i'



σ_i^2=0, $1/\sigma_i^2$= infinity!!

Estimate $1/\sigma_i^2$ using bayes rule

posterior | likelihood | prior

$$P\left(\frac{1}{\sigma_i^2}|data\right) \propto P\left(data|\frac{1}{\sigma_i^2}\right) \times P\left(\frac{1}{\sigma_i^2}\right)$$

Gamma distribution | Gaussian distribution | Gamma distribution

$$\Gamma(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(a)} x^{\alpha-1} e^{-\beta x}$$

$E[\Gamma] = \frac{\alpha}{\beta}$

$Var[\Gamma] = \frac{\alpha}{\beta^2}$

Prior

$E[\Gamma] = \frac{\alpha}{\beta} = \frac{1}{\sigma^2}$ of entire training dataset

Var[Γ] = fitting parameter

Posterior

$$\alpha^* = \alpha + \frac{n}{2}$$

$$\beta^* = \beta + \frac{\sum x_i - \mu}{2}$$

Correct for Dependent Features

- Same training data is used to compute each Z_i and so they may be dependent
- If features are dependent, prior will be underweighted
- Compensate by weighting prior as if features are fully dependent
- Multiply prior weight, $\frac{1}{\tau^2}$, by: $\dfrac{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}{\max(\frac{1}{\sigma_i^2})}$

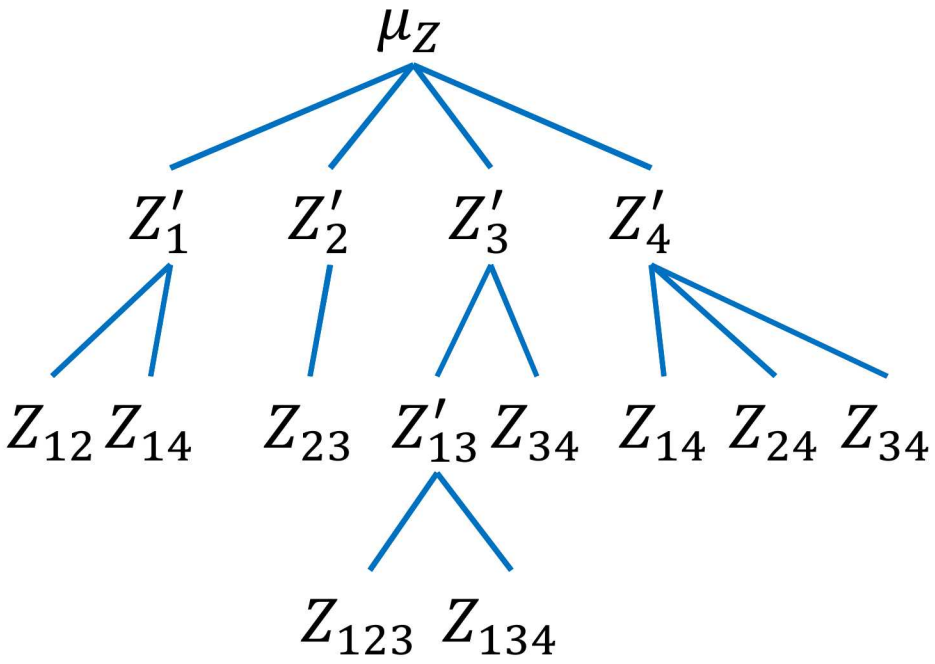
Hierarchically Average the Probability

$$\mu_z = \frac{\sum_i Z_i'/\sigma_i^2 + Z^0/\tau^2}{\sum_i \frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}$$

prior

$$Z_i' = \frac{\sum_{j\neq i} Z_{ij}'/\sigma_{ij}^2 + Z_i/\tau_i^2}{\sum_{j\neq i} \frac{1}{\sigma_{ij}^2} + \frac{1}{\tau_i^2}}$$

$$Z_{ij}' = \frac{\sum_{k\neq j\neq i} Z_{ijk}/\sigma_{ijk}^2 + Z_{ij}/\tau_{ij}^2}{\sum_{k\neq j\neq i} \frac{1}{\sigma_{ijk}^2} + \frac{1}{\tau_{ij}^2}}$$



$Z^0 = < Z >$, Z_i'=estimate of Z_i

$Z_i = < Z | \, x_i = x_i' >$

$Z_{ij} = < Z | \, x_i = x_i', x_j = x_j' >$

$Z_{ijk} = < Z | \, x_i = x_i', x_j = x_j', x_k = x_k' >$

Adding $'$ to Z_i means its an estimate of Z_i given lower levels in the tree

σ and τ have corrections for noise and dependent features respectively

σ is also hierarchically updated

Can Express as a Simple Weighted Average

$$P = \sum_i w_i \times < Z | \, x_i = x_i' > +$$
$$\sum_{ij} w_{ij} \times < Z | \, x_i = x_i', x_j = x_j' > +$$
$$\sum_{ijk} w_{ijk} \times < Z | \, x_i = x_i', x_j = x_j', x_k = x_k' > +$$
$$\ldots$$

Can dynamically update model by updating probabilities. Weights are entirely derived from the probabilities

Fitting the model means optimizing 5 hyperparameters

- nbins: Number of bins for continuous valued data
- TreeDepth, FullyConnectedDepth and FeaturesPerNode
  - Tree creation metaparameters
- σ_{prior}: Variance of prior on sigma

Current Implementation is Slightly Different (Same concepts, derived heuristically)

Sandia National Laboratories

New Model

- Weight probabilities by $\frac{1}{\sigma_i^2}$
- Low data counts:
  - Use prior to specify probability when there are low data counts
  - Use variance of prior as fitting parameter to specify amount of noise in dataset
  - Estimate σ_i^2 based on a prior defined from the entire dataset and likelihood based on training data with feature i=i'
- Class imbalance
  - Adjust variance of prior based on class imbalance to allow rare classes to have larger weights to compensate for lower counts

Old Model

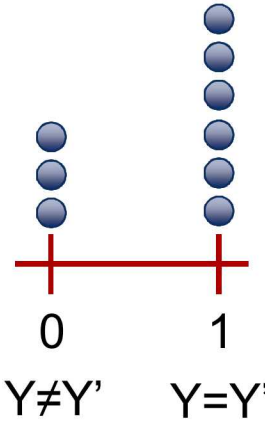
- Weight probabilities by $\frac{1}{\sigma_i^2}$
- Low data counts:
  - limit how large the weight ($\frac{1}{\sigma_i^2}$) can get
  - Add a weight based on data count
  - Arbitrarily define two fitting parameters that define noise
  - Use a "heuristic prior" to specify probability when there are low data counts
- Class imbalance
  - Arbitrarily adjust maximum weight to allow rare classes to have larger weights to compensate for lower counts
- Usefulness / surprise
  - Estimate usefulness by how much each feature changes the probability around the decision function.

Currently Used for Classification

posterior likelihood prior

$$P(\mu_z \mid \cap_i\ Z = Z_i) \propto P(\cap_i\ Z = Z_i \mid \mu_z)\, P(\mu_z)$$

Let the value Z be a probability: $Z_i = P(Y = Y' \mid x_i = x_i')$



Z_i =mean of data

σ_i^2 = variance of data
=variance of Bernoulli distribution
with $P = P(Y = Y' \mid x_i = x_i')$
$= P(Y = Y' \mid x_i = x_i') \times \left(1 - P(Y = Y' \mid x_i = x_i')\right)$

Using same equations (i.e. gaussian distributions) works extremely well!

- Not rigorous, need posterior and likelihood that are both bounded from 0-1
- Want direchlet as likelihood, not prior
- Cannot use multinomial likelihood
  - Evidence needs to be mean of data, not individual data points