Ensemble Learning - 2
Beyond Linear Aggregation

John Klein

centralelille

Goal : Resolve ties on super-majority / approval votes
Use case : Recommender system aggregation

$f_1(\mathbf{x}) =$

1. A
2. C
3. B

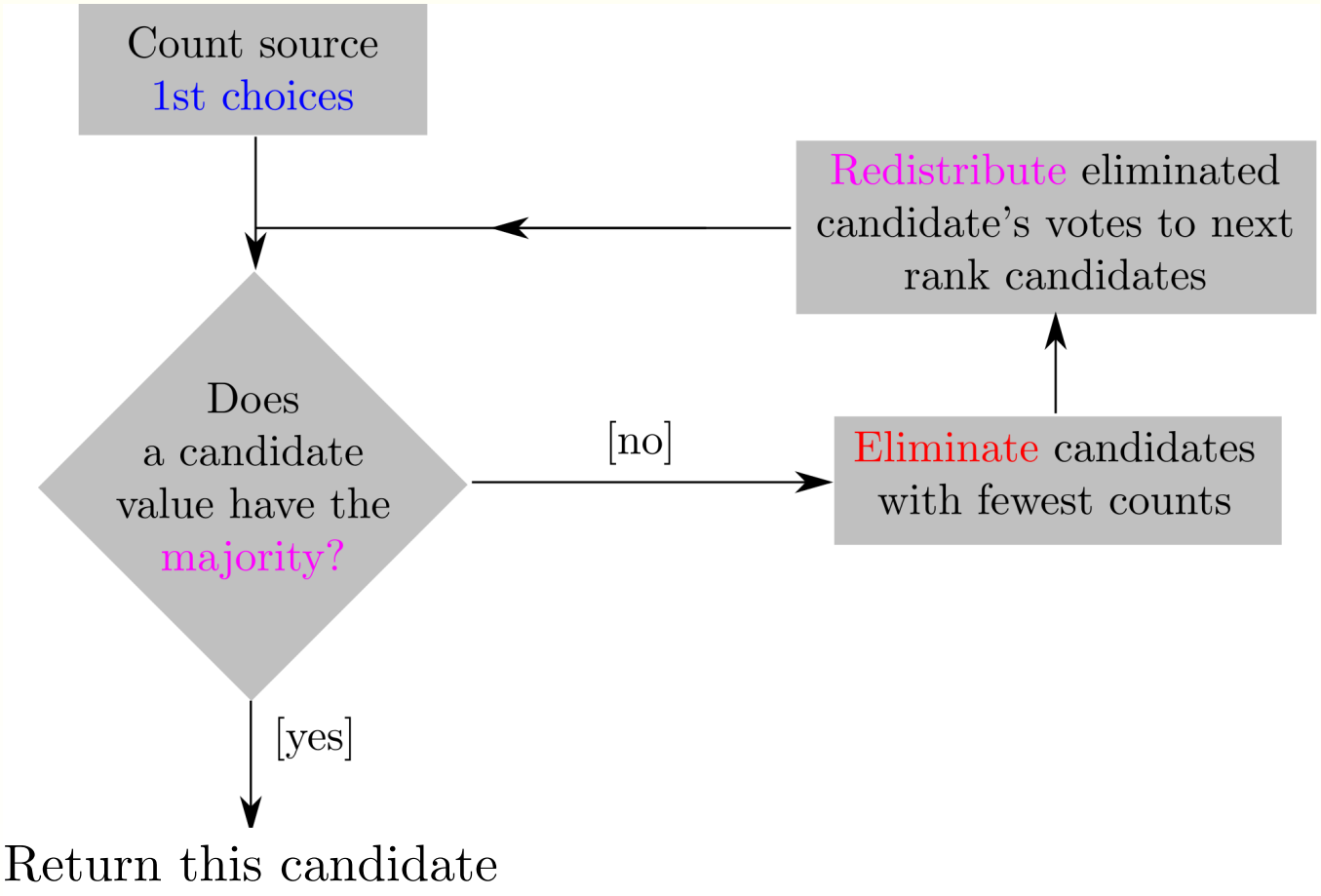
$f_2(\mathbf{x}) =$

1. C
2. B
3. A

$f_{\text{ens}}(\mathbf{x}) =$

1. C
2. A
3. B

Instant Runoff Operator (IRO) :



Example

Fusion of recommender systems

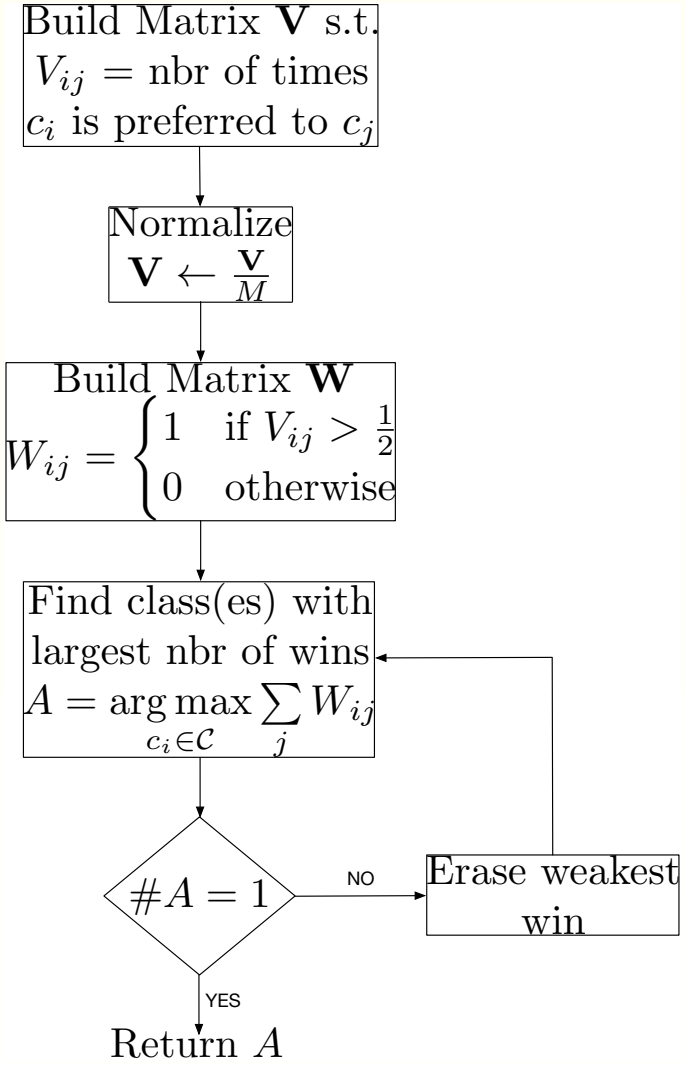
- Suppose we trained M recommender systems each returning a list of $\#\mathcal{C}=4$ objects.
- Each list is ordered from most preferred to least preferred.
- Let us also assume that they delivered only 4 different kinds of ballots.
- The table in the next slide gives the advocacies along with their occurrences.
- Using the majority operator, site web A would win.

Example

Ballot table for example

occurrences	42	26	15	17
ballot	A	B	C	D
	B	C	D	C
	C	D	B	B
	D	A	A	A

Pairwise Rank Operator (PRO) :



Pairwise Rank Operator (PRO) :

- "Weakest" win : find $\arg\min V_{ij}$ among those (i, j) corresponding to a win ($W_{ij} = 1$).
- Let us run PRO on the recommender system fusion example.
- Neither PRO or IRO return a full list of preferred candidates. To obtain the list, eliminate the returned class from the ballots and re-run the algorithm.

Generalized averages :

- F-means are a class of generalized averaging obtained thanks to a bijective mapping $g : \mathcal{Y} \longrightarrow \mathbb{R}$.

$$f_{\text{ens}}(\mathbf{x}) = g^{-1}\left(\frac{1}{M}\sum_{m=1}^{M} g\left(f_m(\mathbf{x})\right)\right).$$

- Typical choices are
  - $g = log$: geometric mean,
  - $g = 1/x$: harmonic mean.

Geometric mean :

$$f_{\text{ens}}(\mathbf{x}) = \left(\prod_{m=1}^{M} f_m(\mathbf{x}) \right)^{\frac{1}{M}} .$$

- Useful when aggregated values are normalized w.r.t some reference value (ratios or percentages).

Example

An orange tree produced in the past years 100, 180, 210 and 300 oranges. The corresponding growth production rates are thus 1.8, 1.17 and 1.44. The arithmetic mean of these rates is 1.47 whereas the geometric mean is 1.44. Starting with 100 oranges and applying a growth rate of 1.47 gives 314.47 oranges while using 1.44 as growth rate gives exactly 300 oranges.

Harmonic mean :

$$f_{\text{ens}}(\mathbf{x}) = \frac{M}{\sum_{m=1}^{M} \frac{1}{f_m(\mathbf{x})}}.$$

- Useful when aggregated values are unnormalized growth rates.
- Harmonic mean $<$ Geometric mean $<$ Arithmetic mean.

Example

A car is moving by a 100m at 50km/h and by another 100m at 70km/h. The car displacement duration is 7.2 for the first move followed by 5.14s for the second move which makes 12.34s in total.

- Arithmetic mean of $(50; 70)$ is 60. The displacement duration obtained by moving at 60km/h on 200m is 12s.
- Geometric mean of $(50; 70)$ is 59.16. The displacement duration obtained by moving at 59.16km/h on 200m is 12.17s.
- Harmonic mean of $(50; 70)$ is 58.33. The displacement duration obtained by moving at 58.33km/h on 200m is exactly 12.34s.

Non-linear Aggregation

Classification : combining predictions that are class label conditional **probabilities**

Aggregation of probability distributions : desirable properties

- Let p_{ens} denote the aggregated distribution.

- Axiom (i) : weak set wise function property (WSFP)

Definition (WSFP)

For all subset $A \subseteq \mathcal{C}$,

$$p_{\text{ens}}(A) = g^{(A)}(p_1(A), \ldots, p_M(A)), \tag{1}$$

for some function $g^{(A)} : [0;1]^M \to [0;1]$.

- Interpretation : the aggregated chances of event A are depending solely on the input probabilities on the same event A.

Aggregation of probability distributions : desirable properties

- Axiom (ii) : strong set wise function property (SSFP)

Definition (SSFP)

For all subset $A \subseteq \mathcal{X}$,

$$p_{\text{ens}}(A) = g(p_1(A), \ldots, p_M(A)), \tag{2}$$

for some function $g : [0;1]^M \to [0;1]$.

- Interpretation : same as before but the combination rule is the same for each event otherwise relabeling the elements of $\mathcal{C}$ would impact the fusion.

Aggregation of probability distributions : desirable properties

- Axiom (iii) : unanimity (or idempotence)

Definition (Unanimity)

If $p_m = p_0$ for all m, then $p_{\text{ens}} = p_0$.

- Interpretation : if the sources are unanimous, then the aggregate distribution is a copy of the input ones.

Proposition

If $\#\mathcal{C} \geq 3$, a probability distribution aggregation operator satisfying SSFP and unanimity is an LOP (convex combination).

Aggregation of probability distributions : desirable properties

- Axiom (iv) : independence preservation (IP)

Definition (IP)

For any two subsets A and B of $\mathcal{C}$ s.t.
$p_m(A \cap B) = p_m(A) \times p_m(B) \ \forall m$, then
$p_{\text{ens}}(A \cap B) = p_{\text{ens}}(A) \times p_{\text{ens}}(B)$

- Interpretation : when input distributions are unanimously independent w.r.t. a pair of events, this is also true for the aggregated distribution.

- No LOP operator achieves IP except if $w_m = 1$ for some m (dictatorship or selection).

Non-linear Aggregation

Classification : combining predictions that are class label conditional **probabilities**

Aggregation of probability distributions : desirable properties

- In the Bayesian setting, the **posterior** p' is an update of the **prior** p through the **likelihood** function L : $p'(x) \propto L(\mathcal{D})\, p(x)$.
- Axiom (v) : Bayesian externality (EB)

Definition (EB)

Let $(p')_{\text{ens}}$ denote the combination of the updated distribution p'_m using likelihood function L and $(p_{\text{ens}})'$ denote the updated combination of the distributions p_m using the same likelihood function. Then

$$(p')_{\text{ens}} = (p_{\text{ens}})'.$$

- Interpretation : Bayesian update and fusion commute.
- The time at which some information arrives does not matter.

Aggregation of probability distributions : desirable properties

Proposition

If a probability distribution fusion operator writes

$$p_{ens} = \frac{1}{Z} \prod_{m=1}^{M} (p_m)^{w_m} ,$$

where coefficients w_m are non-negative and sum to one : $\sum_{m=1}^{M} w_m = 1$ and Z is a normalization constant, then it achieves unanimity and EB.

- These operators are known as **logarithmic opinion pool** (LogOP) operators.
- LogOP = weighted geometric mean !

Aggregation of probability distributions : desirable properties

- Another nice property of LogOPs :

Proposition

An aggregated distribution returned by a LogOP is the solution of the minimization problem :

$$\arg\min_{p} \sum_{m=1}^{M} w_m d_{KL}\left(p, p_m\right),$$

where d_{KL} is the Kullback-Leibler divergence.

- LogOPs are "barycenters" in the KL divergence sense.

Non-linear Aggregation

Classification : combining predictions that are class label conditional **probabilities**

Aggregation of probability distributions : probabilistic calculus

- Each distribution p_m can be interpreted as $p_Y\left(.|\mathbf{x}, z_m\right)$
- Z_m is a random variable attached to the mth predictor, ex : $Z_m = \mathcal{D}^{(m)}$, private dataset of the mth learner.
- For simplicity, we assume $M = 2$.
- From Bayesian standpoint, we want to infer $p_Y\left(.|\mathbf{x}, \mathcal{D}^{(1)}, \mathcal{D}^{(2)}\right)$.
- Let $w_m = p\left(\mathcal{D}^{(m)}|\mathbf{x}\right)$.

Non-linear Aggregation

Classification : combining predictions that are class label conditional **probabilities**

Aggregation of probability distributions : probabilistic calculus

- Applying Bayes, we can write

$$p\left(y|\mathbf{x},\mathcal{D}^{(1)},\mathcal{D}^{(2)}\right) \propto p\left(\mathcal{D}^{(1)},\mathcal{D}^{(2)}|y,\mathbf{x}\right) \times p_0\left(y|\mathbf{x}\right) \tag{3}$$

- Assuming conditional independence, we have

$$p\left(y|\mathbf{x},\mathcal{D}^{(1)},\mathcal{D}^{(2)}\right) \propto p\left(\mathcal{D}^{(1)}|y,\mathbf{x}\right) \times p\left(\mathcal{D}^{(2)}|y,\mathbf{x}\right) \times p_0\left(y|\mathbf{x}\right) \tag{4}$$

- Applying Bayes again gives

$$p\left(y|\mathbf{x},\mathcal{D}^{(1)},\mathcal{D}^{(2)}\right) \propto \frac{p\left(y|\mathcal{D}^{(1)},\mathbf{x}\right) w_1}{p_0\left(y|\mathbf{x}\right)} \times \frac{p\left(y|\mathcal{D}^{(2)},\mathbf{x}\right) w_2}{p_0\left(y|\mathbf{x}\right)} \times p_0\left(y|\mathbf{x}\right) \tag{5}$$

Non-linear Aggregation

Classification : combining predictions that are class label conditional **probabilities**

Aggregation of probability distributions : probabilistic calculus

- This also writes

$$p\left(y|\mathbf{x},\mathcal{D}^{(1)},\mathcal{D}^{(2)}\right) \propto \frac{p\left(y|\mathcal{D}^{(1)},\mathbf{x}\right) w_1 \times p\left(y|\mathcal{D}^{(2)},\mathbf{x}\right) w_2}{p_0\left(y|\mathbf{x}\right)} \quad (6)$$

- $p_0\left(y|\mathbf{x}\right)$ is a prior $\neq p\left(y|\mathbf{x}\right)$.
- General case :

$$p\left(y|\mathbf{x},\mathcal{D}^{(1)},..,\mathcal{D}^{(M)}\right) \propto \frac{\prod_{m=1}^{M} p\left(y|\mathcal{D}^{(m)},\mathbf{x}\right) w_m}{\left(p_0\left(y|\mathbf{x}\right)\right)^{M-1}}. \quad (7)$$

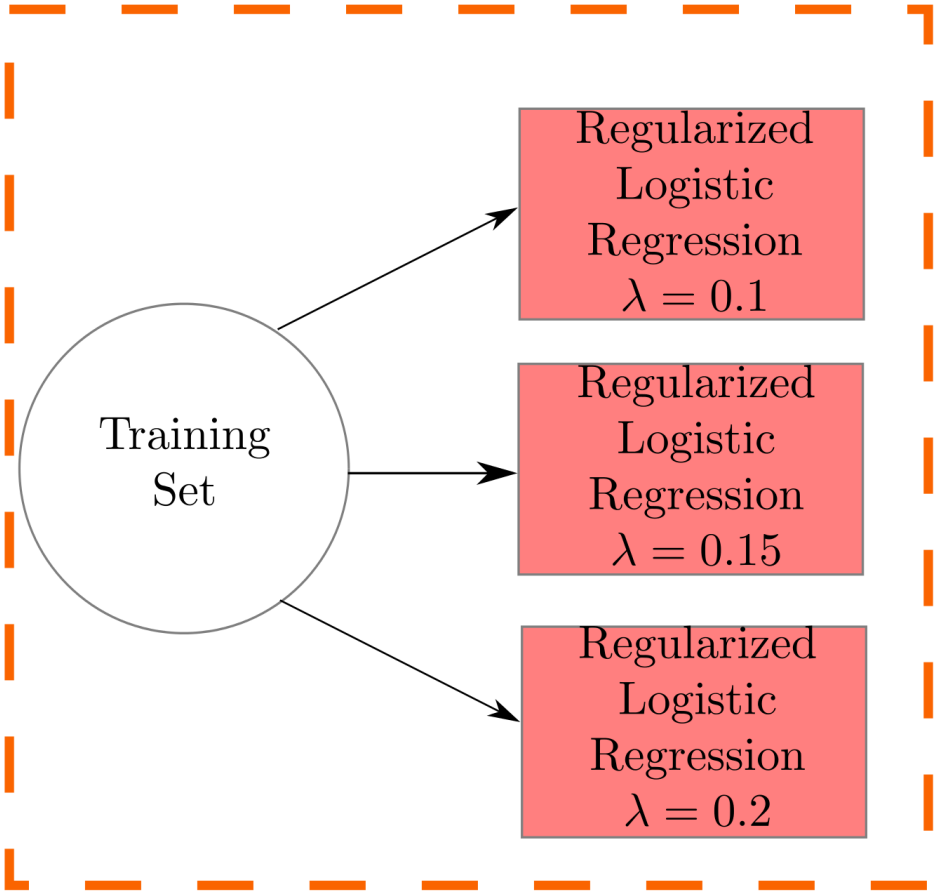
Non-linear Aggregation

Classification : combining predictions that are class label conditional **probabilities**

Aggregation of probability distributions : when they are no longer defined on the same σ-algebra.

- Suppose f_1 is a trained logistic regressor to discriminate c_1 from $B_1 = \{c_2, .., c_\ell\}$
- The σ-algebra on which the distribution attached to f_1 is defined is : $\{\emptyset, \{c_1\}, B_1, \mathcal{C}\}$.
- If each f_m is a one-versus-all probabilistic classifier, their distributions cannot be aggregated using the aforementioned techniques.
- Generalized frameworks can help : imprecise probabilities, belief functions/random sets.

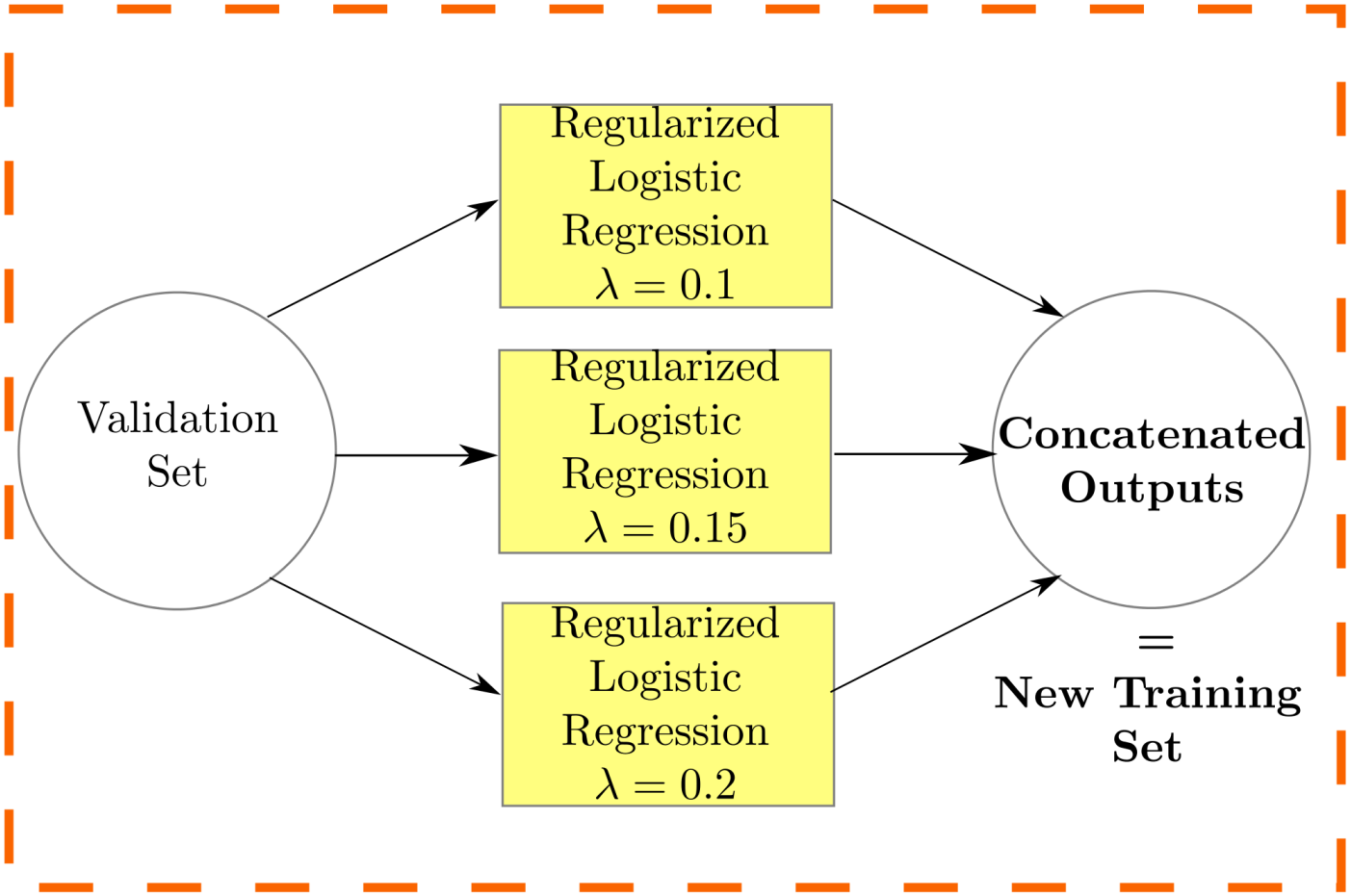
Stacking : Procedure



Step 1

Stacking : Procedure

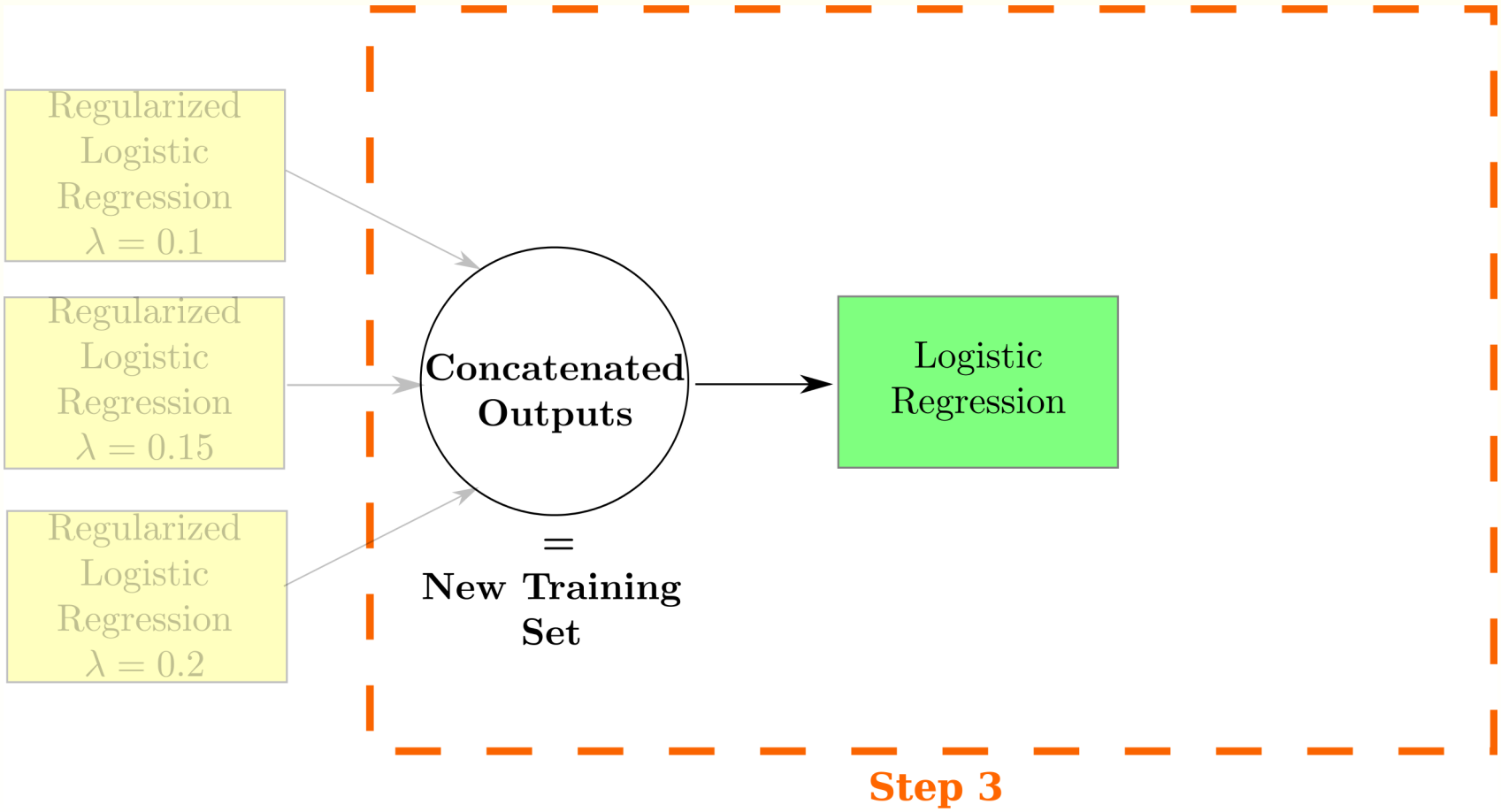
Generate a new training set for a second-stage classifier.



Step 2

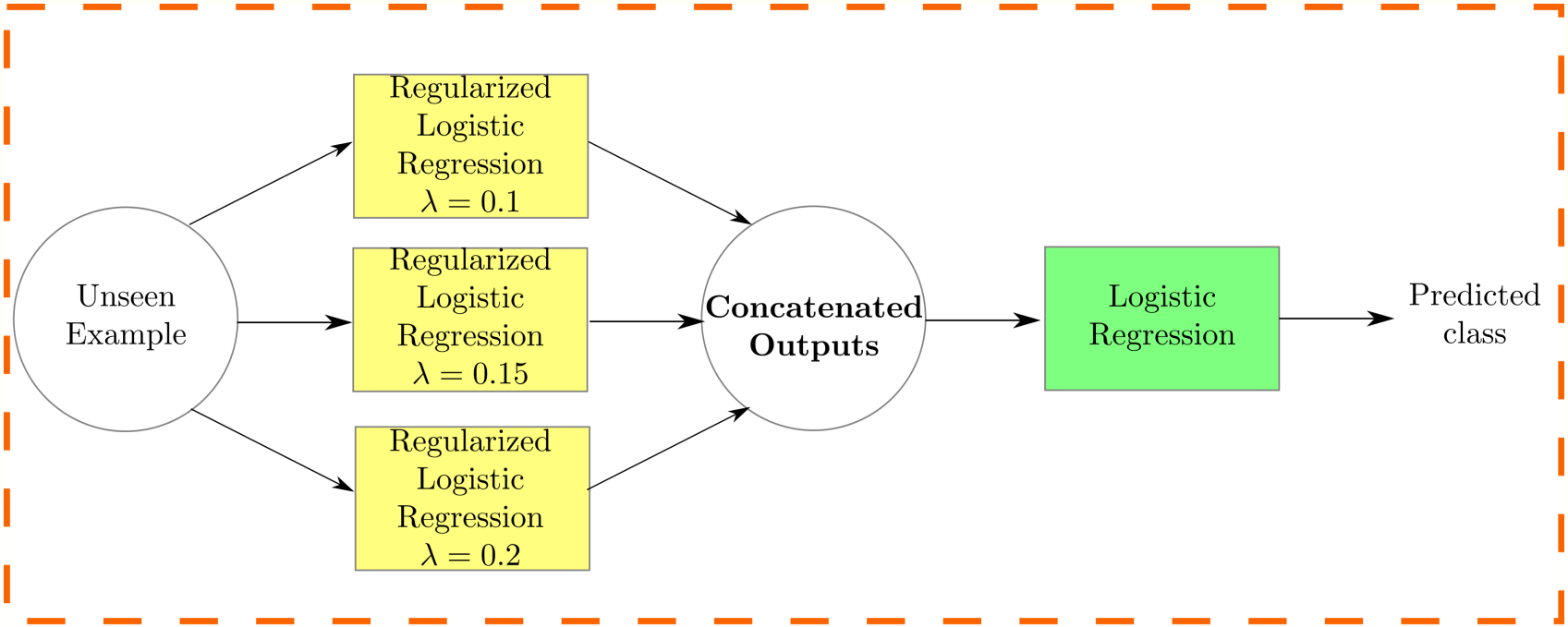
Stacking : Procedure

Train the second-stage classifier.



Stacking : Procedure

Use it all.



Step 4

Stacking : Comments

- Stacking can be used for heterogeneous classifiers too.
- Need for validation data for second-stage training.
- Solution : repeat the procedure in a **cross validation** (CV) fashion.
- Leave-one-out CV (**LOOCV**) : let $f_m^{(-i)}$ denote the mth predictor trained on the whole training set **except** data point $\mathbf{x}^{(i)}$, then in step 3, one needs to minimize

$$J(\mathbf{w}) = \sum_{i=1}^{n} L\left(y^{(i)}, \mathrm{sgm}\left(\sum_{m=1}^{M} w_m f_m^{(-i)}\left(\mathbf{x}^{(i)}\right)\right)\right).$$

Any optimal solution ?

- We assume $f_m(\mathbf{x}) \in \mathcal{C}$ (predictions are class labels).
- Let $\mathbf{c} = [f_1(\mathbf{x}) \dots f_M(\mathbf{x})]$.
- Based on the sole information contained in vector $\mathbf{c}$, and under 0-1 loss, the optimal decision is

$$f_{\text{ens}}(\mathbf{x}) = \arg\max_y p(y|\mathbf{c}(\mathbf{x})).$$

- Can we infer the conditional distributions $p(y|\mathbf{c}(\mathbf{x}))$?
- Applying Bayes, we have $p(y|\mathbf{c}) \propto p(\mathbf{c}|y)\, p(y)$.
- $p(\mathbf{c}|y)$ is a joint distribution whose marginals are multinomial.
- How many parameters do we have to learn ?

Any optimal solution ?

- Not scalable : polynomial memory complexity w.r.t. $\#\mathcal{C}$ and exponential w.r.t. M !
- Conditional independence assumptions :

$$p\left(\mathbf{c}|y\right) = \prod_{m=1}^{M} p\left(c_m|y\right).$$

- Linear memory complexity w.r.t. both $\#\mathcal{C}$ and M.
- This is stacking, with Naive Bayes classifier as second-stage classifier !