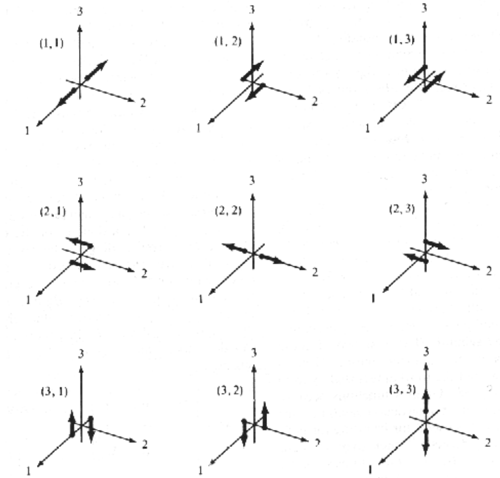
Machine Learning in Geophysics

Lecture 3 – The Machine Learning Workflow

From previous lectures

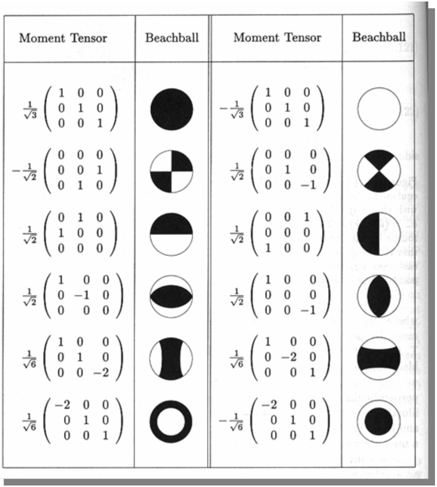
Moment tensor components



Point sources can be described by the seismic moment tensor M. The elements of M have clear physical meaning as forces acting on particular planes.

From previous lectures

Beachballs and moment tensor



Moment Tensor	Beachball	Moment Tensor	Beachball	
$\frac{1}{\sqrt{3}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$		$-\frac{1}{\sqrt{3}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$		explosion - implosion
$-\frac{1}{\sqrt{2}}\begin{pmatrix}0&0&0\\0&0&1\\0&1&0\end{pmatrix}$		$\frac{1}{\sqrt{2}}\begin{pmatrix}0&0&0\\0&1&0\\0&0&-1\end{pmatrix}$		vertical strike slip fault
$\frac{1}{\sqrt{2}}\begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix}$		$\frac{1}{\sqrt{2}}\begin{pmatrix}0&0&1\\0&0&0\\1&0&0\end{pmatrix}$		vertical dip slip fault
$\frac{1}{\sqrt{2}}\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix}$		$\frac{1}{\sqrt{2}}\begin{pmatrix}1&0&0\\0&0&0\\0&0&-1\end{pmatrix}$		45° dip thrust fault
$\frac{1}{\sqrt{6}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&-2\end{pmatrix}$		$\frac{1}{\sqrt{6}}\begin{pmatrix}1&0&0\\0&-2&0\\0&0&1\end{pmatrix}$		
$\frac{1}{\sqrt{6}}\begin{pmatrix}-2&0&0\\0&1&0\\0&0&1\end{pmatrix}$		$-\frac{1}{\sqrt{6}}\begin{pmatrix}-2&0&0\\0&1&0\\0&0&1\end{pmatrix}$		compensated linear vector dipoles

Standard procedure

1. Determine moment tensor

$$\mathbf{M} = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{pmatrix}$$

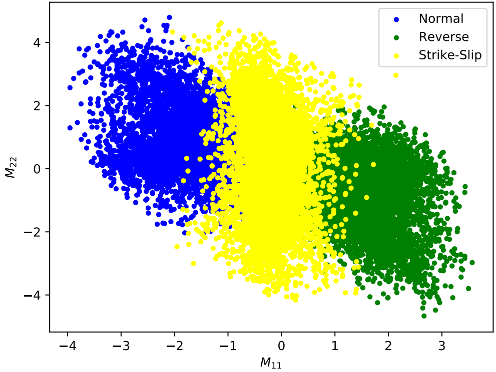
2. Determine eigenvectors (principal directions)
3. Make beach ball diagram
4. Interpret mechanism (strike-slip, normal, reverse)

We will use machine learning to go from 1 directly to 4.

Supervised machine learning strategy

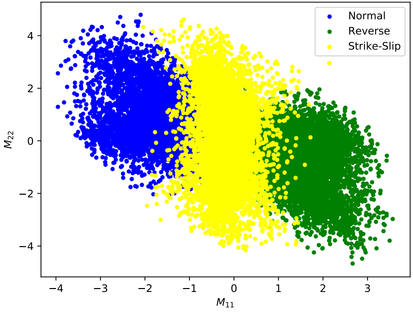
1. Get a dataset with known solutions
2. Inspect dataset to identify problems/useful features
3. Prepare dataset for machine learning algorithm (normalization, categorization)
4. Split data into training data, validation data, test data.
5. Train ML on training data
6. Use validation data to tweak parameters
7. Check on test data
8. Make predictions on data with unknown solutions

The training data



Two moment tensor elements M_{11}, M_{22}, fault category (color)

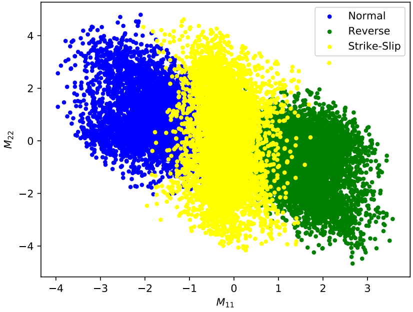
The training data



Question

What issues will we have if we want to apply a SVM to this problem?

The training data



Question

What issues will we have if we want to apply a SVM to this problem?

Answers

- Have 3 categories, not a binary classification problem
- Categories not separated by straight lines

Data preparation

- Need to normalize input data to either 0 mean and standard deviation 1, or to range $0.0 \ldots 1.0$
- Need to turn labels into numbers

Here we have 6 labels *'N'*, *'N-SS'*, *'R'*, *'R-SS'*, *'SS'*, *'SS-N'*, *'SS-R'*, want only 3 *'N'*, *'R'*, *'SS'* and need to transform to numbers $\{0, 1, 2\}$.

Splitting the data

Split the labelled dataset into three parts

Training data: Used to train the ML algorithm, ca 80% of total data

Validation data: Check accuracy of ML algorithm after each training (different parameters), ca 10% of total data

Test data: Should only be used once for checking prediction accuracy after all adjustments have been made, ca 10% of total data

Considerations

Question

What do we need to consider when splitting the data?

Considerations

Question

What do we need to consider when splitting the data?

Answers

- Each dataset must be representative, i.e. contain samples from all categories.
- Cannot have duplicates between different sets.
- Need to have enough data in each set.

Training - validation workflow

1. Choose hyperparameters, e.g. regularization parameter C
2. Fit model to training data
3. Evaluate prediction on validation data
4. Go back to 1 until optimal model found

Need some evaluation criteria.

Confusion matrix

For classification, we can construct the *Confusion Matrix*

prediction outcome

		A	**B**	**total**
actual value	$\mathbf{A'}$	True **A**	False **B**	A'
	$\mathbf{B'}$	False **A**	True **B**	B'
total		A	B	

Precision and recall

If we write the confusion matrix

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

Then we can calculate precision

$$P_i = \frac{C_{ii}}{\sum_j C_{ji}} \text{ e.g. } P_1 = \frac{C_{11}}{C_{11} + C_{21}}$$

and recall

$$R_i = \frac{C_{ii}}{\sum_j C_{ij}} \text{ e.g. } R_1 = \frac{C_{11}}{C_{11} + C_{12}}$$

Final test

Once all adjustments have been made, we can then perform the final test on the test data.

Question

Why can't we just use the validation data, as the it has not been used in the fitting?

Final test

Once all adjustments have been made, we can then perform the final test on the test data.

Question

Why can't we just use the validation data, as the it has not been used in the fitting?

Answers

- Have adjusted hyper-parameters to optimize performance.
- Specifics of data can influence our choices
- Not a true reflection of performance on completely new data

Multi-class classification with SVM

We can extend a binary classifier, e.g. SVM, to multi-class classification with a *one-vs-rest* Strategy.

- One classifier for each class $A, B, C, \ldots$
- Each classifier is trained A vs. not-A, B vs. not-B, ...
- Each classifier provides score, e.g. distance from decision boundary
- Class with highest score wins

Summary

- Goal of supervised ML workflow is to tune parameters for optimal prediction
- Essential aspect, division of available data in training, validation, test
- Test dataset can only be used once all adjustments have been made
- Confusion matrix, precision, recall are three ways to judge prediction performance