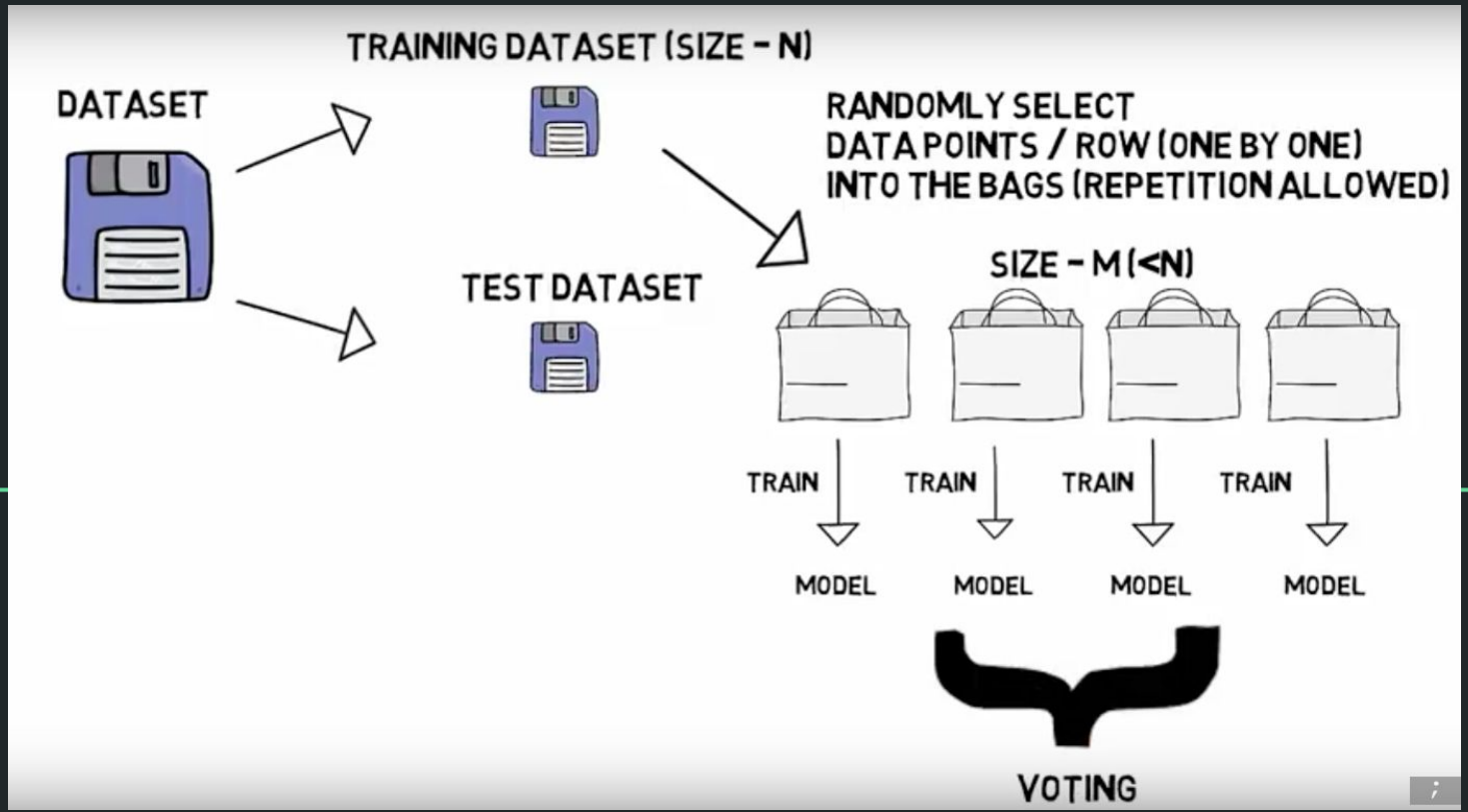
Error metrics & Cross-validation

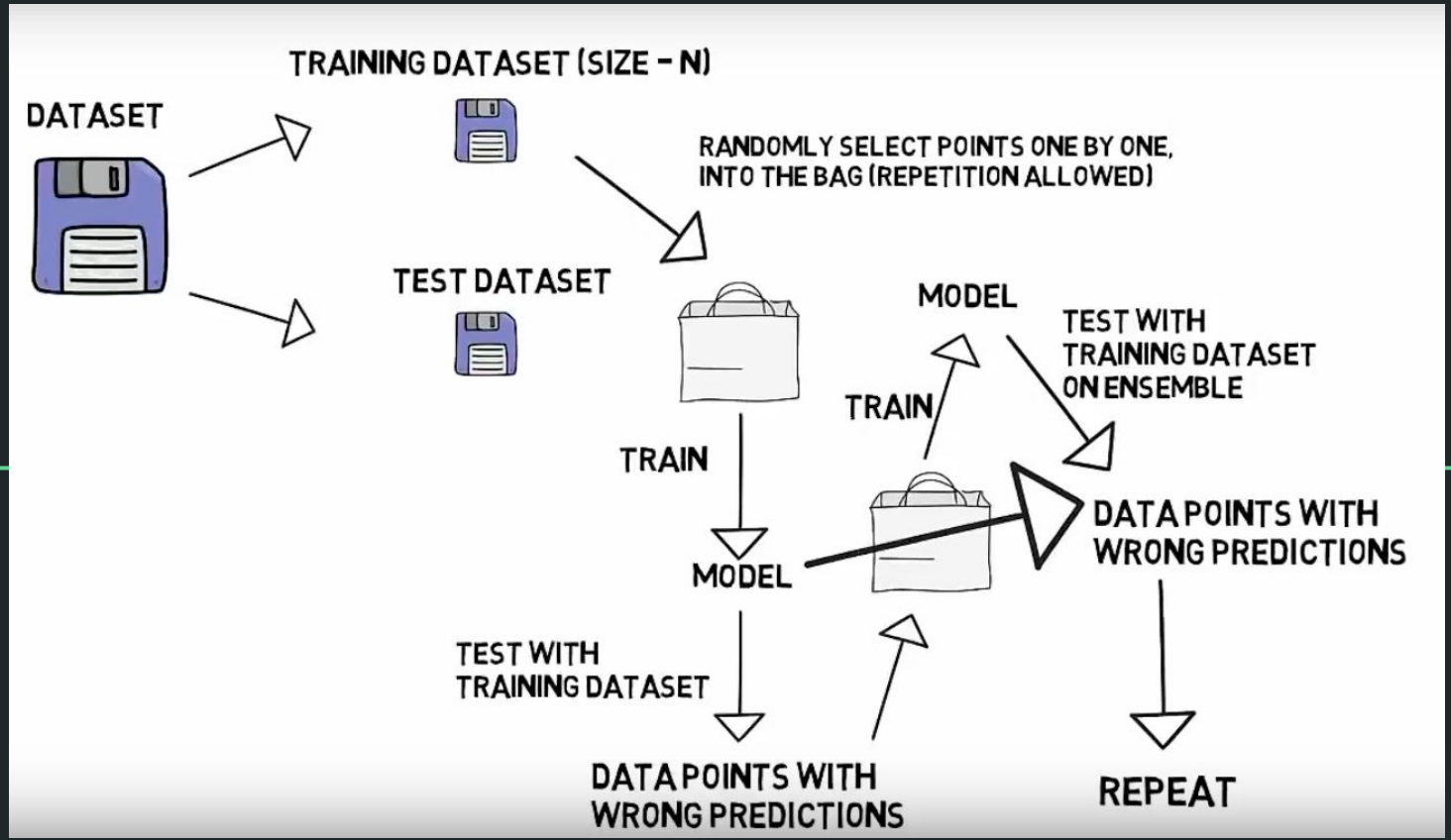
Data Mining & Analytics

(adapted from INFO 254 by Zach Pardos)

Ensemble: Bagging

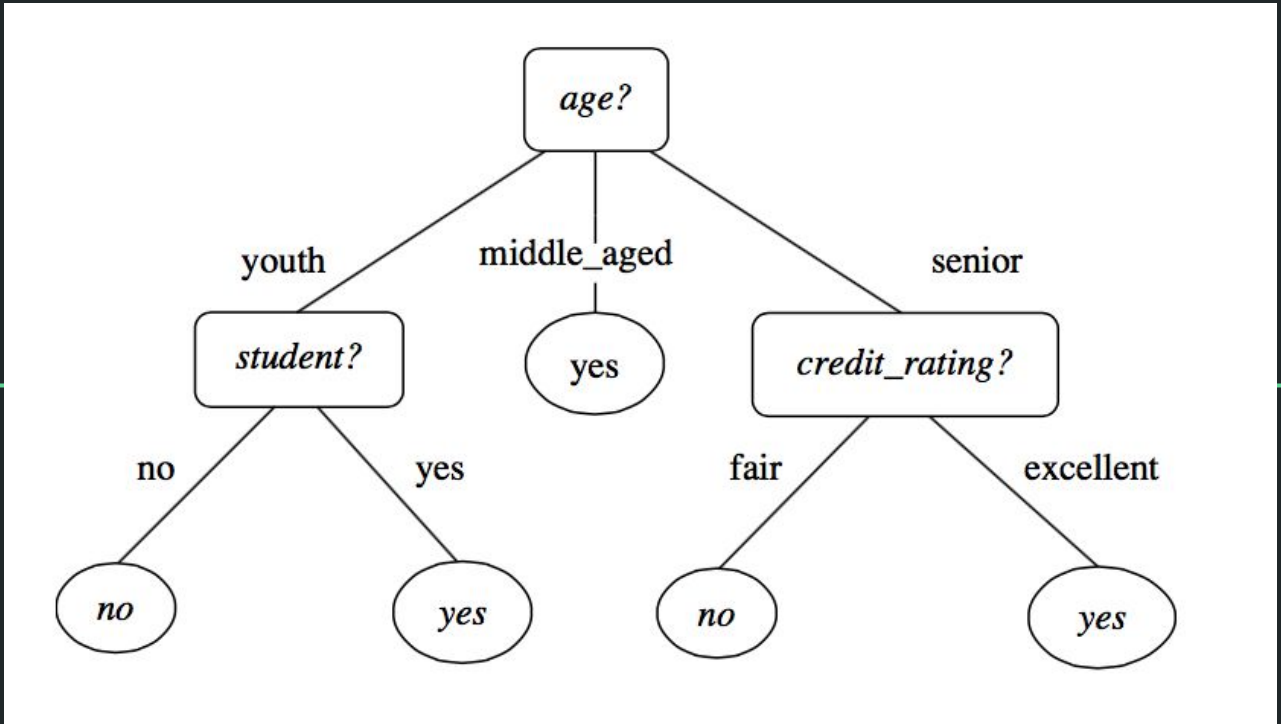


Ensemble: Boosting

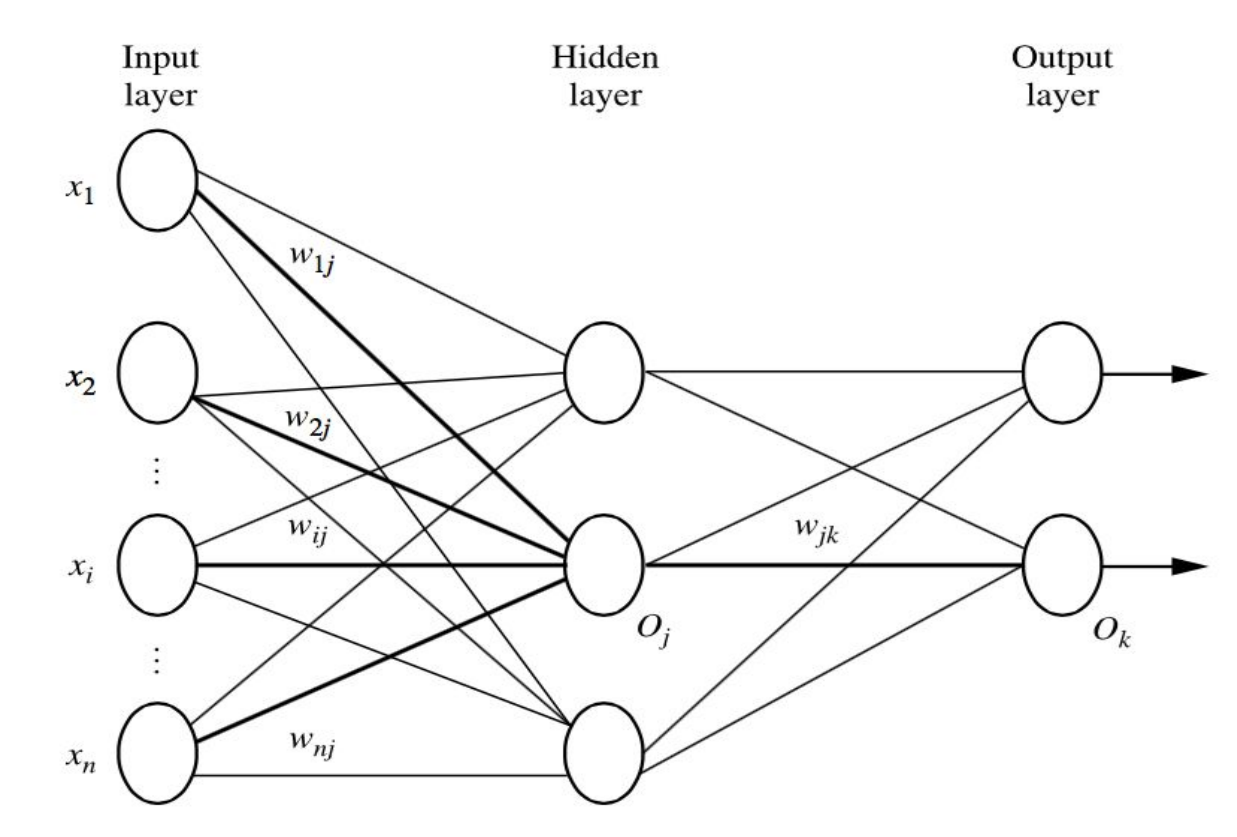


Error metrics: Motivation

What metrics were used to obtain the best decision tree split?



What was the goodness metric used to train the model?



Error Metrics Vs Loss Functions

Those were model *loss functions*

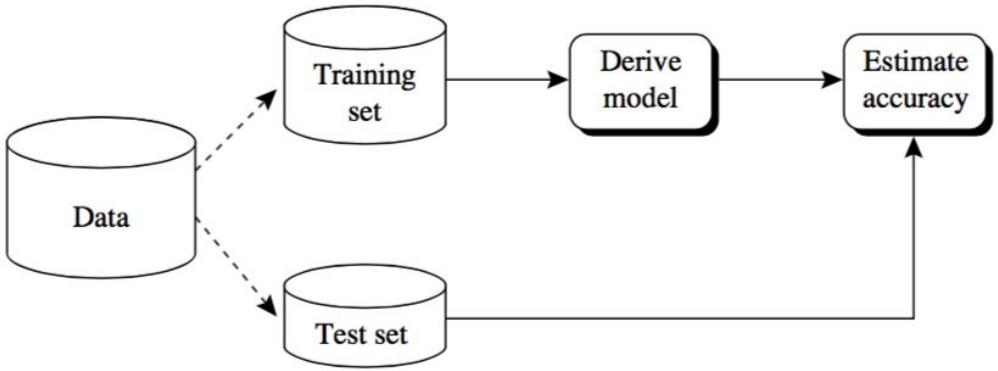
- Designed to provide the best chance for model convergence
- Intended to correlate with model fit to data

Error metrics, on the other hand, are:

- Meant to be proxies for outcomes you care about (in the domain)
- Selection of error metric is application dependent
- Can incorporated into training to inform the stopping criterion

Motivation

- How do you define the “goodness” of your model?
  - Its ability to generalize to unseen data
  - What can affect generalization goodness?
    - Choice of
      - Algorithm
      - Hyper parameter values
      - Training data
      - Random initializations (random seed can be seen as a hyper parameter)
    - Each set of choices (i.e. models) can result in different generalization goodness



Motivation

- How do you define the “goodness” of your model?
  - Its ability to generalize to unseen data
  - What can affect generalization goodness?
    - Choice of
      - Algorithm
      - Hyper parameter values
      - Training data
      - Random initializations (random seed can be seen as a hyper parameter)
    - Each set of choices (i.e. models) can result in different generalization goodness
  - Generalization goodness can be quantified using *error metrics*:
    - accuracy, precision, recall
    - mean absolute error (MSE)
    - root mean squared error (MSE)
    - area under the ROC curve (AUC)

Error Metrics for Classification

Confusion matrix

		Predicted class: *yes*	Predicted class: *no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

Confusion matrix

- **True positives** (*TP*): These refer to the positive tuples that were correctly labeled by the classifier. Let *TP* be the number of true positives.
- **True negatives** (*TN*): These are the negative tuples that were correctly labeled by the classifier. Let *TN* be the number of true negatives.
- **False positives** (*FP*): These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class *buys_computer* = *no* for which the classifier predicted *buys_computer* = *yes*). Let *FP* be the number of false positives.
- **False negatives** (*FN*): These are the positive tuples that were mislabeled as negative (e.g., tuples of class *buys_computer* = *yes* for which the classifier predicted *buys_computer* = *no*). Let *FN* be the number of false negatives.

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Which rows are:

- TP, FP, TN, FN?

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Accuracy = ?

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Accuracy = 60%

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Error = ?

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Error = 40%

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Sensitivity = ?
(aka recall)

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Sensitivity = $66.\overline{66}\%$
(aka recall)

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Specificity = ?

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Specificity= 50%

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Precision = ?

Han, Kamber, Pei (2011), Sec 8.5

Error Metrics for Classification

These metrics apply when a prediction is being made of a class/category. These metrics also apply when the label is binary and the prediction is rounded to 0 or 1.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

predicted	actual
0	0
0	1
1	0
1	1
1	1

Precision = $66.\overline{6}\%$

Han, Kamber, Pei (2011), Sec 8.5

In-class Exercise

Choose 3 of the following scenarios and for each scenario, come up with a metric appropriate for judging the quality of predictions.

1) Predicting black friday category in Kaggle class competition
2) Predicting lung cancer from chest x-rays
3) Predicting high-school GPA
4) Evaluating search engine results
5) Predicting the location of an object in 3D space
6) Predicting if a Twitter user is a liberal or conservative

Is the prediction categorical or continuous? What aspects of the prediction problem might inform your metric consideration?

Error Metrics for Regression (numeric)

These metrics apply when a prediction is being made of a numeric target/label, including when the label is binary and the prediction is continuous between 0 & 1

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

predicted	actual
0.25	0
0.45	1
0.66	0
0.71	1
0.70	1

Error Metrics for Regression (numeric)

These metrics apply when a prediction is being made of a numeric target/label, including when the label is binary and the prediction is continuous between 0 & 1

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

predicted	actual
0.25	0
0.45	1
0.66	0
0.71	1
0.70	1

MAE = ?

Error Metrics for Regression (numeric)

These metrics apply when a prediction is being made of a numeric target/label, including when the label is binary and the prediction is continuous between 0 & 1

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

predicted	actual
0.25	0
0.45	1
0.66	0
0.71	1
0.70	1

MAE = 0.41

Error Metrics for Regression (numeric)

These metrics apply when a prediction is being made of a numeric target/label, including when the label is binary and the prediction is continuous between 0 & 1

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

predicted	actual
0.25	0
0.45	1
0.66	0
0.71	1
0.70	1

RMSE = ?

Error Metrics for Regression (numeric)

These metrics apply when a prediction is being made of a numeric target/label, including when the label is binary and the prediction is continuous between 0 & 1

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

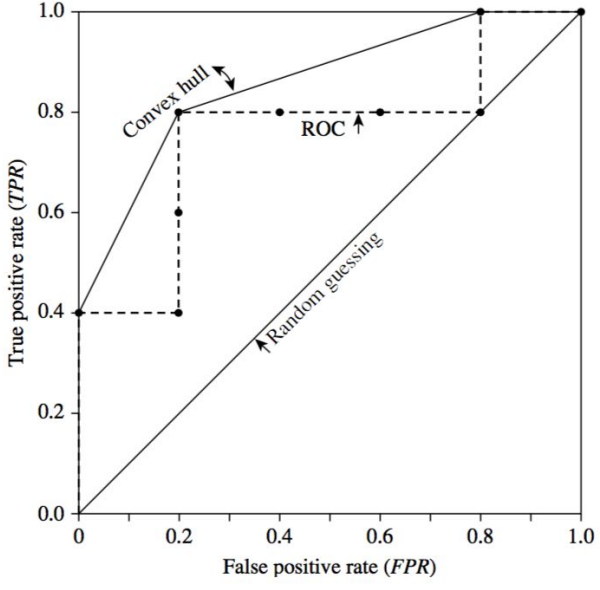
predicted	actual
0.25	0
0.45	1
0.66	0
0.71	1
0.70	1

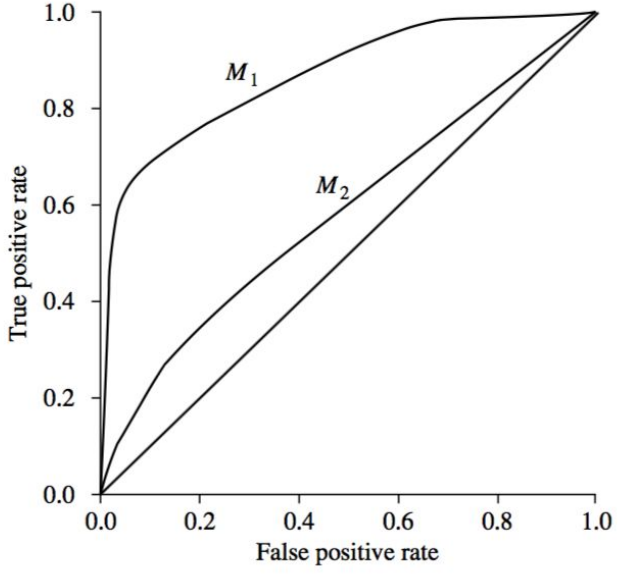
RMSE =~ 0.4415

RMSE accentuates the impact of outliers

Area Under the [RO] Curve (AUC)

Receiver Operator





Exact calculation: Probability that a randomly chosen positive label prediction will be greater than a randomly chosen negative label prediction
AUC = 0.50 means no better than random chance

Han, Kamber, Pei (2011), Sec 8.5

AOC Explained: https://youtu.be/OAI6eAyP-yo?t=88

Visualization: http://www.navan.name/roc/

Motivation: Cross Validation

How do you know that the goodness of your model is reliable?

Cross-validation example

Step 1: choose a K (typically 5 or 10) and randomly assign each row to a single fold (between 1 and K). This fold assignment indicates the fold in which that row will serve as a tuple in the test set

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	0.40	0.32	0.55	0.65
Aurora	1	0.40	0.76	0.85	0.91
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	0.40	0.60	0.54	0.67
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	0.40	0.32	0.55	0.65

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	0.40	0.32	0.55	0.65
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	incorrect	incorrect	incorrect	incorrect
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 2: Conduct K phases of training/testing, starting with fold 1, which will serve as the test set, with the other folds serving as the training set.

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	incorrect	correct	correct	correct
Palo	5	0.40	0.32	0.25	0.15
Sophia	2	correct	incorrect	incorrect	incorrect
Miguel	1	incorrect	correct	correct	correct
Aurora	1	correct	correct	correct	correct
Jess	3	incorrect	correct	incorrect	correct

Cross-validation example

Step 3: Calculate error metric on the whole dataset (concatenation of results from each fold)

Student	Fold	Item c1	Item c2	Item d1	Item d2
Bob	4	0.40	0.32	0.55	0.65
Palo5	5	0.40	0.32	0.25	0.15
Sophia	2	0.40	0.60	0.54	0.67
Miguel	1	0.40	0.32	0.55	0.65
Aurora	1	0.40	0.76	0.85	0.91
Jess	3	0.40	0.32	0.55	0.65

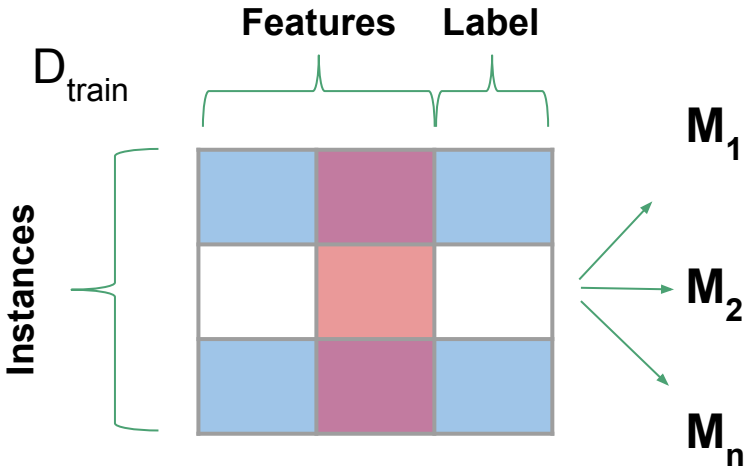
Cross-validation

- **k-fold:** tuples randomly partitioned into *k* mutually exclusive subsets
- **Leave-one-out:** k = number of tuples
- **Stratified:**class distribution of the tuples in each fold is approximately the same as that in the initial data

Ensemble Methods

Primary ensemble methods: blending, bagging, boosting, stacking & blending

Blending: Make predictions on a test and validation set. Train a blending model based on the validation set. Apply blending model to the test set.



Predictions $F(M_{1...n}) \rightarrow Y$ **Label**

$D_{validate}$

M_1	M_2	M_n		
0.95	0.95	0.95		1
0.95	0.50	0.95		1
0.05	0.05	0.50		0
0.50	0.95	0.95		1

D_{test}

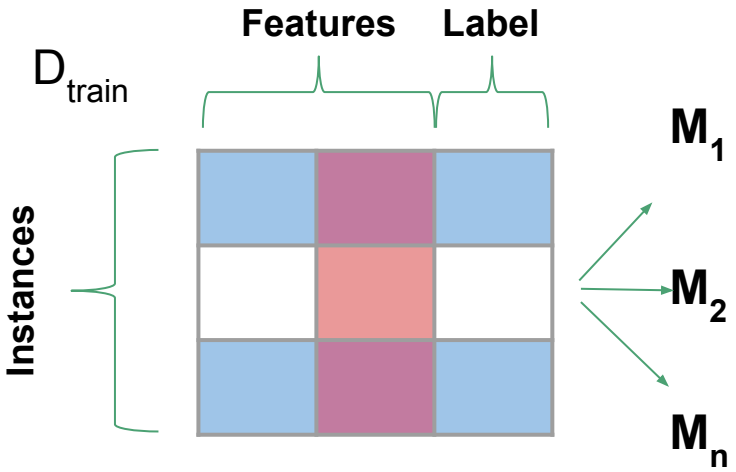
0.41	0.51	0.27		0
0.77	0.11	0.89		1

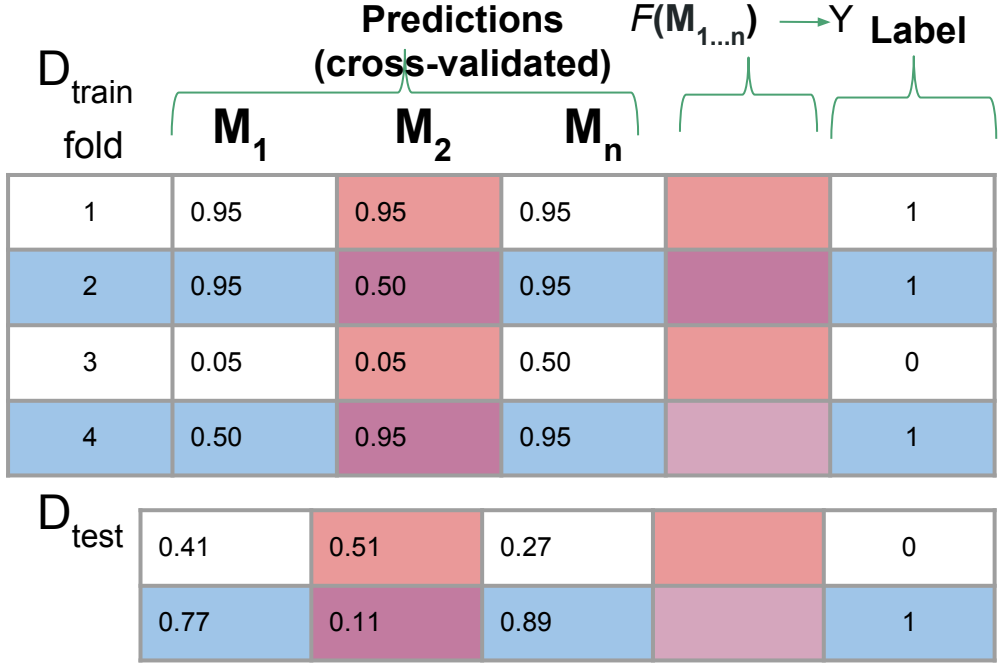
Ensemble Methods

Primary ensemble methods: blending, bagging, boosting, stacking & blending

Stacking: same as blending but uses *cross-validation*

Train and predict using an ensemble of models using cross-validation on the training set.
Train those same models on the entire training set, make predictions on test set.
Train a combiner on the cross-validated training set.
Apply trained combiner to the test set.





In-class Exercise

1) Predicting lung cancer from chest x-rays (recall, precision, Fq)
2) Predicting high-school GPA (MAE)
3) Evaluating search engine results (Recall)
4) Predicting the location of an object in 3D space (Euclidean / cosine distance)
5) Predicting if a Twitter user is a liberal or conservative (AUC)

Is the prediction categorical or continuous? What aspects of the prediction problem might inform your metric consideration?

Classification metrics: Accuracy, Error, Precision, Recall, F1, AUC*
Regression metrics: MAE, RMSE, AUC (binary)

*AUC requires probabilities associated with a binary class prediction

Midterm

- Thursday, March 14th (10 days)
- Two sided page of notes allowed (printed or handwritten)
- Otherwise closed computer/book
- There will be extra credit
- Questions ranging from pseudo-code to conceptual covering lectures, labs, and readings
- Sit at least 2 chairs away from lab partner(s)