Stat 215A - Week 10

Zoe Vernon
Thanks to Rebecca Barter for sharing her slides

Upcoming schedule for the class

Today (10/26):

- ❏ Group project (lab 4) assigned. Due **Friday 11/9**.
- ❏ Evaluating classification algorithms.

Next Friday (10/2):

- ❏ Causal inference

Friday (11/9): Lab 4 due

Tuesday (11/13): midterm

Friday (11/16): Final project assigned. Due **12/7** (3 weeks).

Evaluating classification algorithms

Evaluating classification algorithms

There are many ways to evaluate the accuracy of classification

- Test/validation sets
- Cross-validation
- Confusion matrix
- TP, FP, TN, FN
- ROC curve
- Confidence interval for error

Train-validation-test sets

Train-validation-test sets

Split your data into 3 parts

1. Training set
2. Validation set
3. Test set

The size of splits depends on the algorithm you are training.

Train-validation-test sets

Split your data into 3 parts

1. Training set: train the model on this data
   a. e.g. run K-nearest neighbors on your data
2. Validation set: data used for tuning hyperparameters (don't "learn" from this data
   a. e.g. tune choice choice of K (by evaluating models trained with different K's on the validation set)
3. Test set: only used once as a test of your final model

Cross validation

Cross validation is a way to doing validation while only splitting your data into train and test sets.

Do K-fold cross validation on the training data to evaluate the models and tune hyperparameters.

After selecting a model, use the test set to check how your model performs.

Confusion matrix

Confusion matrix

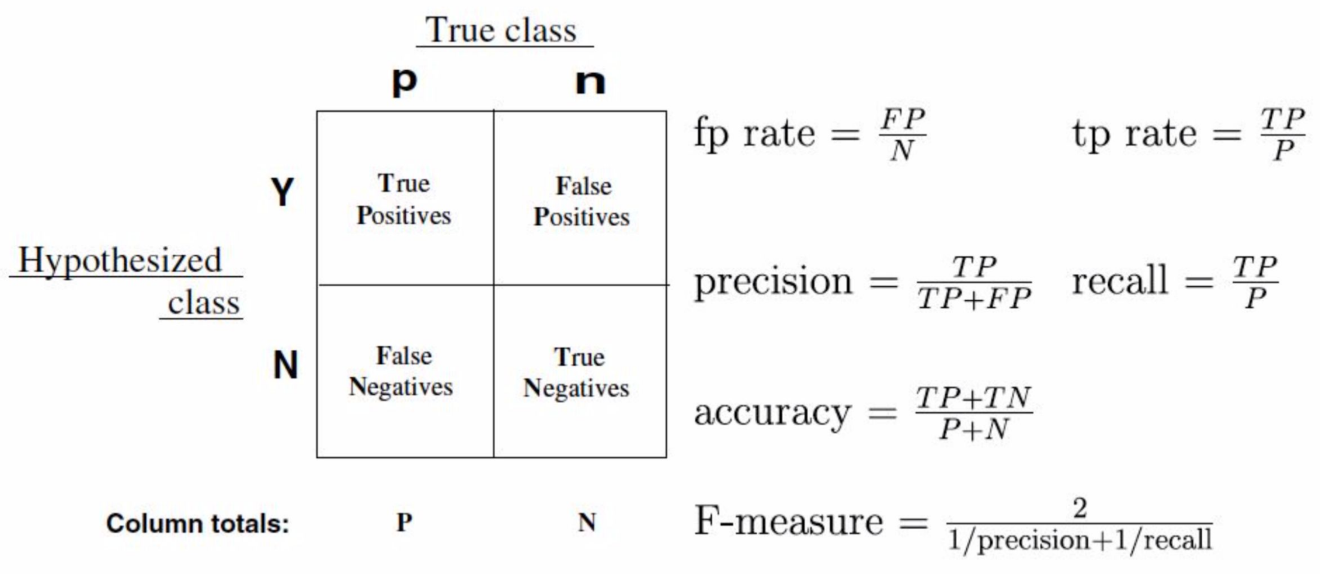


Fig. 1. Confusion matrix and common performance metrics calculated from it.

Source: Fawcett (2005)

Confusion matrix

		True class: p	True class: n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

ROC curve

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

Fig. 1. Confusion matrix and common performance metrics calculated from it.

Source: Fawcett (2005)

Confusion matrix

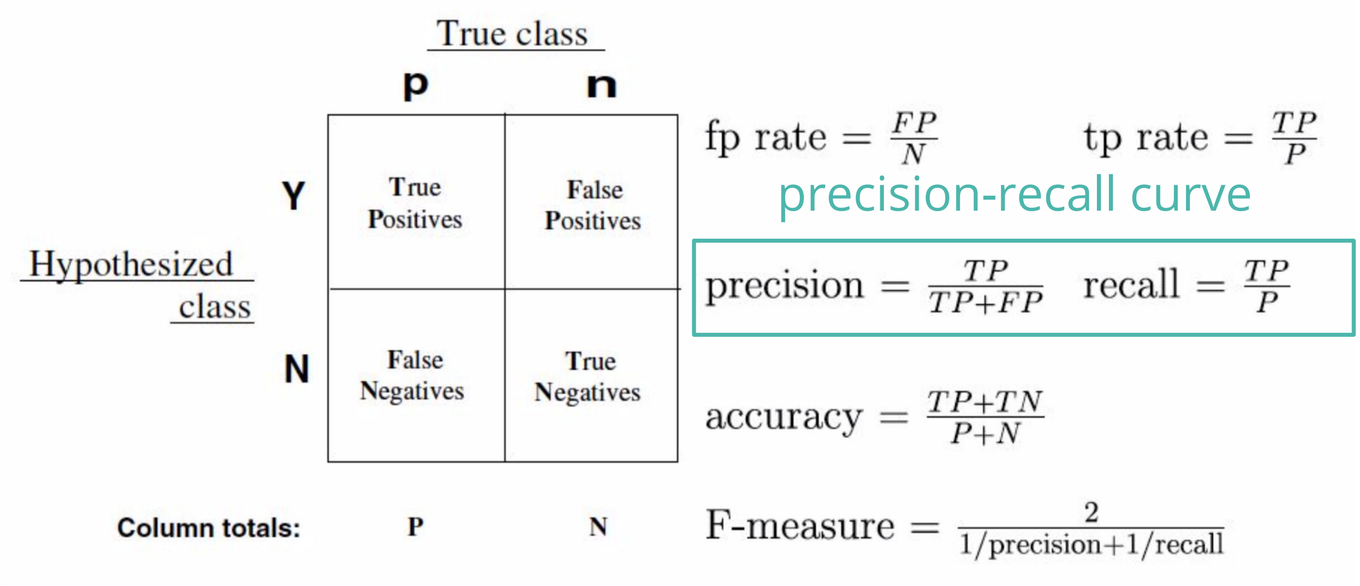


Fig. 1. Confusion matrix and common performance metrics calculated from it.

Source: Fawcett (2005)

Receiver operating characteristics (ROC) graphs

Receiver operating characteristics (ROC) graphs

Used to depict the tradeoff between “hit rates” and false alarm rates” of classifiers

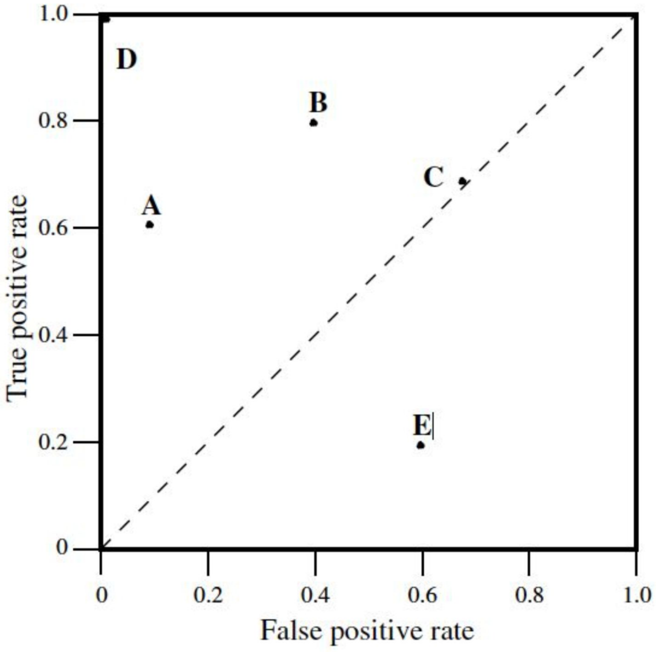


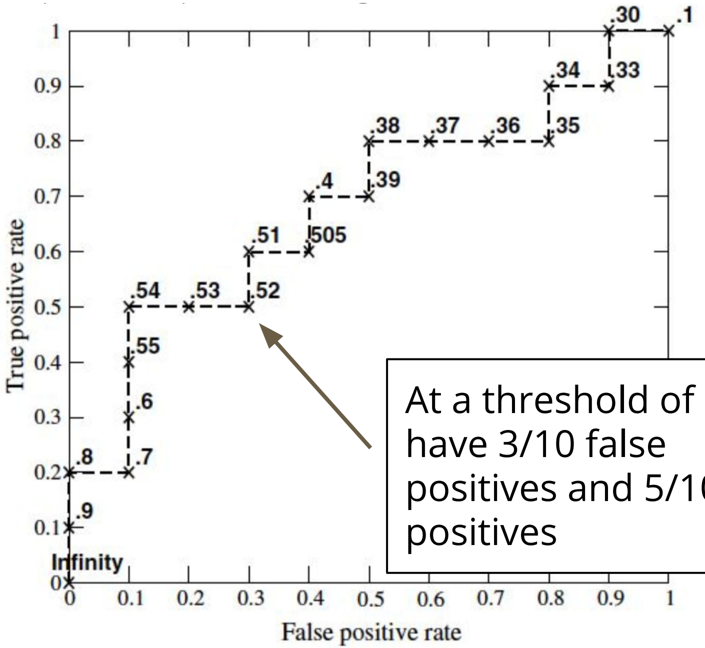
Fig. 2. A basic ROC graph showing five discrete classifiers.

Source: Fawcett (2005)

Receiver operating characteristics (ROC) graphs

We can generate an ROC curve when the output of a classifier is a probability and we must choose a threshold for the final predicted class

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



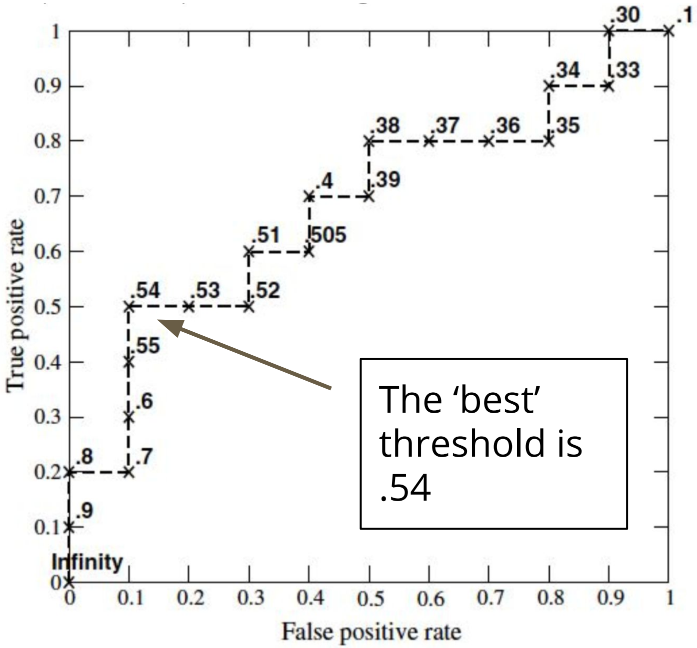
At a threshold of .52 we have 3/10 false positives and 5/10 true positives

Source: Fawcett (2005)

Receiver operating characteristics (ROC) graphs

We can generate an ROC curve when the output of a classifier is a probability and we must choose a threshold for the final predicted class

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



Source: Fawcett (2005)

Receiver operating characteristics (ROC) graphs

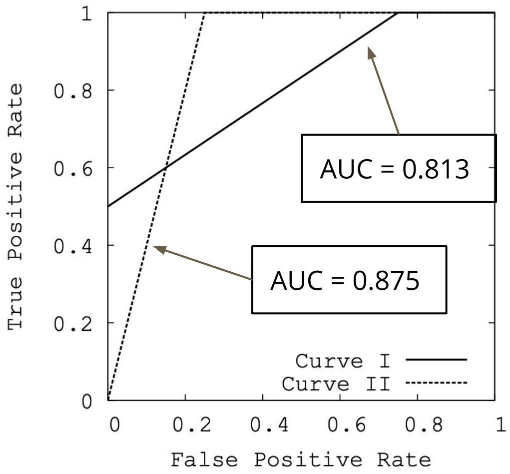
Question:

Would this method give an accurate reflection when there is class imbalance?

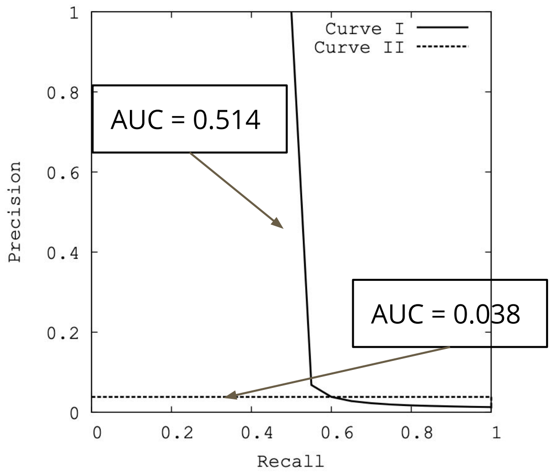
(e.g. 10% of the of the observations are in group 1 and 90% are in group 2)

Receiver operating characteristics (ROC) graphs

Question: Would this method give an accurate reflection when there is class imbalance?



(a) Comparing AUC-ROC for two algorithms



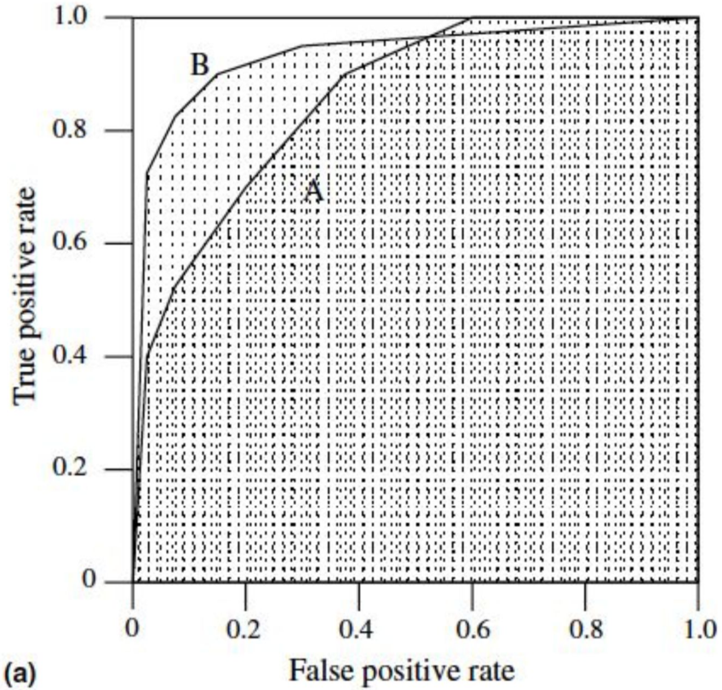
(b) Comparing AUC-PR for two algorithms

Figure 7. Difference in optimizing area under the curve in each space

Source: Davis and Goadrich (2006)

The area under the ROC curve (AUC)

The area under the curve (AUC) is a method for comparing algorithms



(a)

Source: Fawcett (2005)

The area under the ROC curve (AUC)

The area under the curve (AUC) is a method for comparing algorithms and evaluating classifiers.

The AUC has an important statistical property:

> *The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance*

Source: Fawcett (2005)

The area under the ROC curve (AUC)

Care should be taken when using ROC curves to compare classifiers

- The ROC graph is often used to select the best classifiers simply by graphing them in ROC space and seeing which one dominates.

Source: Fawcett (2005)

The area under the ROC curve (AUC)

Care should be taken when using ROC curves to compare classifiers

- ❏ The ROC graph is often used to select the best classifiers simply by graphing them in ROC space and seeing which one dominates.
- ❏ This is misleading: it is analogous to taking the maximum of a set of accuracy figures from a single test set.

Source: Fawcett (2005)

The area under the ROC curve (AUC)

Care should be taken when using ROC curves to compare classifiers

- The ROC graph is often used to select the best classifiers simply by graphing them in ROC space and seeing which one dominates.
- This is misleading: it is analogous to taking the maximum of a set of accuracy figures from a single test set.
- Without a measure of **variance** we cannot compare classifiers

Source: Fawcett (2005)

The area under the ROC curve (AUC)

Care should be taken when using ROC curves to compare classifiers

- ❑ The ROC graph is often used to select the best classifiers simply by graphing them in ROC space and seeing which one dominates.
- ❑ This is misleading: it is analogous to taking the maximum of a set of accuracy figures from a single test set.
- ❑ Without a measure of **variance** we cannot compare classifiers

It is a good idea to the average of multiple ROC curves (e.g. via cross validation)

See Fawcett (2005) for examples on how to average

Source: Fawcett (2005)

The area under the ROC curve (AUC)

Example: classify.R

Lab 4: cloud detection

Groups: see lab4_groups.csv

Name	Group	Name	Group	Name	Group
Linqing Wei	1	Aummul Baneen	5	Xiao Yun Chang	9
Sonali Dayal	1	Yuchen Zhang	5	Armin Askari	10
Michael Lim	1	Rui Chen	6	Ang Li	10
Jiajian Lu	2	Benji Lu	6	Taejoo Ahn	10
Kehsin Su	2	David Chen	6	Zihao Chen	11
Marius Wiggert	2	Tiffany Tang	7	Todd Faulkenbeery	11
Gaowei Chen	3	Danni Deng	7	Yutong Wang	11
Alan Aw	3	Bassel Sadek	7	Adelson Chua	11
Baishan Guo	3	Hari Das	8	Dan Soriano	12
Nicholas Sim	4	James Duncan	8	Kevin Benac	12
Fan Dong	4	Michelle Yu	8	Qi Chen	12
Mengling Liu	4	Alon Amid	9	Anran Hu	13
Lei Zhang	5	Qi Chen	9	Shuni Li	13
				Philippe Boileau	13

Data: satellite images over arctic region

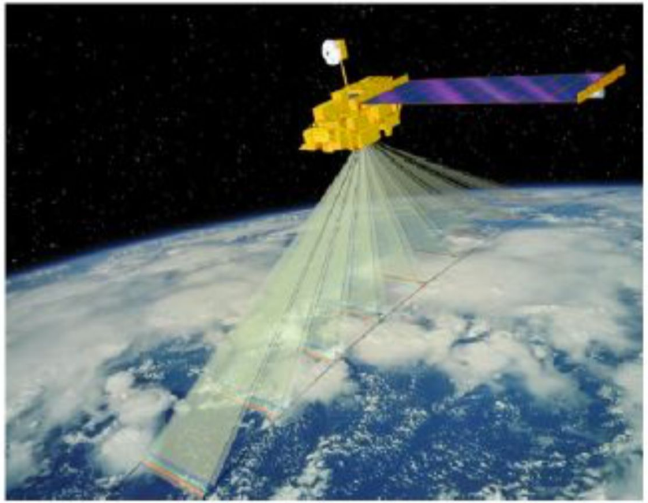


Figure 1. Cartoon illustration of the Terra satellite with the view directions of the nine MISR cameras. Image is courtesy of the MISR science team at the Jet Propulsion Laboratory.

Data: satellite images over arctic region

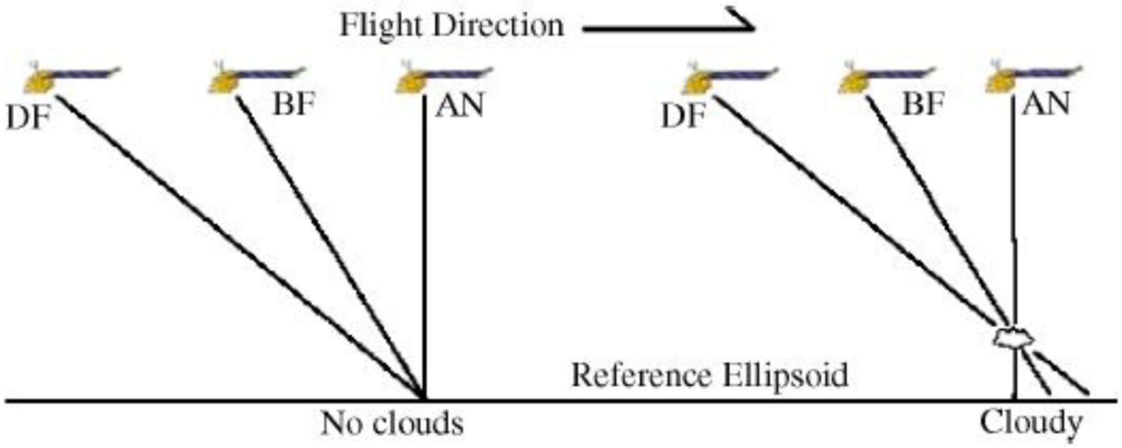


Figure 2. Registration of surface features and clouds to the reference ellipsoid. Note that only three of the nine MISR cameras are illustrated and that surface objects are registered to the same location, whereas clouds are registered to different locations.

Data: satellite images over arctic region

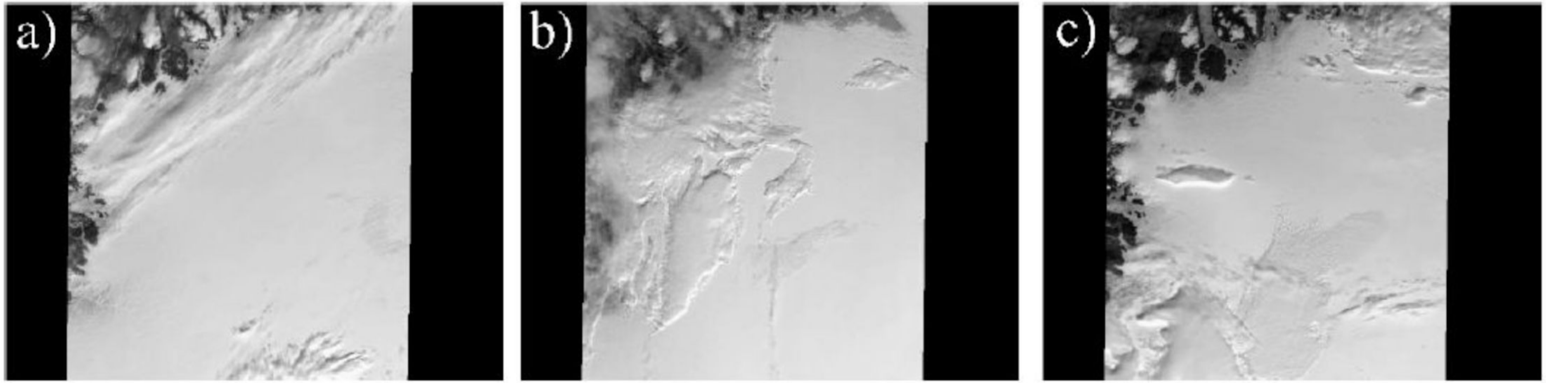


Figure 3. Data collected by the MISR An-camera for three consecutive orbits (i.e., 13257, 13490, and 13723) over blocks 20–22 of path 26.

Data: satellite images over arctic region

Figure 4. Expert labels for blocks 20–22 of MISR orbits (a) 13257, (b) 13490, and (c) 13723. White represents high confidence cloudy; gray, high confidence clear; and black, unlabeled pixels.

Lab 4 demo: cloud_data.R