STAT 215A Fall 2019 Week 11

Tiffany Tang

11/8/19

Announcements

- No GSI office hours this Monday due to Veteran’s Day
- Lab 4 (group project) due in two weeks: Thursday, November 21 at 11:59pm
- Good job on the midterm!
  - Median: 34/38

Midterm T/F

1. When you calculate the ordinary least squares estimator, the residual vector always has mean zero.

2. Suppose that $\hat{\mu}$ is an estimator of some parameter of interest μ. Then the (population) MSE of $\hat{\mu}$ is given by

$$\mathbb{E}[(\hat{\mu} - \mu)^2] = Bias^2(\mu) + Var(\mu)$$

8. Under the linear regression model $y_i = x_i^\top \beta + \epsilon_i$, the errors ϵ_i must be i.i.d. normally distributed and have mean 0 in order for the OLS estimator $\hat{\beta}_{OLS}$ to be an unbiased estimator of β.

Function Documentation

- What does this function do?
- Describe the inputs and outputs
- Like a mini R help page

```
CalculateSampleCovariance <- function(x, y, verbose = TRUE) {
  # Computes the sample covariance between two vectors.
  # Args:
  #   x: One of two vectors whose sample covariance is to be calculated.
  #   y: The other vector. x and y must have the same length, greater than one,
  #      with no missing values.
  #   verbose: If TRUE, prints sample covariance; if not, not. Default is TRUE.
  # Returns:
  #   The sample covariance between x and y.
...
}
```

Plan for Today

- Crash course in classification algorithms
  - Logistic Regression
  - Naive Bayes
  - Discriminant Analysis
  - KNN Classifier
- Next time
  - Maximum margin classifiers/SVMs
  - Random Forests
  - Ensembles
  - Evaluation metrics

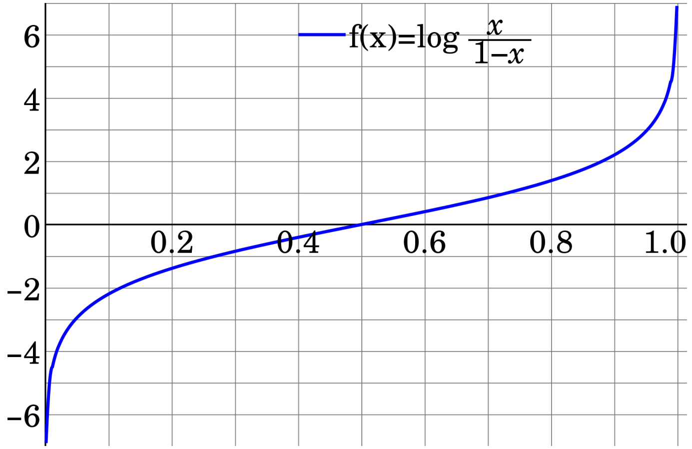
Why classification and not regression?

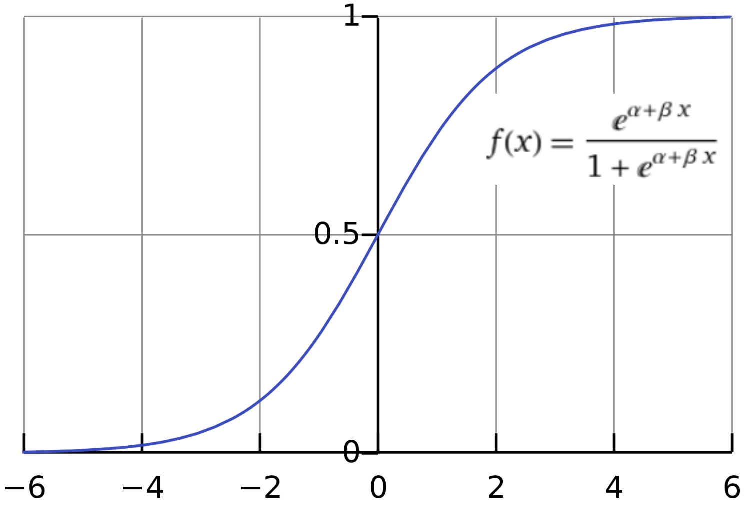
- Suppose we have data $X_1, \ldots, X_n$ and responses $y_1, \ldots, y_n$, but the responses are categorical (i.e., $y_i \in \{1, \ldots, K\}$)
- Problems with regression:
  - Hard to assign numeric values to categories
  - Usually no ordering of the categories
  - Even if categories are ordered, not necessarily equally spaced

Logistic Regression

- Assume there are two classes and $y_i \mid x_i \sim Bern(p_i)$ are independent with

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i \qquad \Longleftrightarrow \qquad p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$





- Solve MLE via Newton-Rhapson or iteratively reweighted LS

- Can either output the fitted probabilities $\widehat{p_i}$ or round them to the *most likely* class (i.e., class 0 or class 1)

Logistic Regression Extensions

- What if we have more than 2 classes?
  - Multinomial logistic regression
- What if we have p > n (or simply p is large)?
  - Regularized logistic regression

$$max_{\alpha,\beta}\ \ell(\alpha,\beta,X) - \ \lambda P(\beta)$$

- Something to think about carefully: why the logistic model and not some other model?

Naïve Bayes

- Central quantity of interest in classification: P(Y = k | X)
  - That is, given data X, what is the probability that it is in class k
  - Decision rule: if we knew P(Y = k | X) for each k, predict the class with the highest probability
- Idea: use Bayes rule to estimate P(Y = k | X)

$$P(Y = k \mid X) = \frac{P(X \mid Y = k)\, P(Y = k)}{P(X)} \propto \underbrace{P(X \mid Y = k)}_{\text{likelihood}}\, \underbrace{P(Y = k)}_{\text{prior}}$$

- Define $P(Y = k) = \pi_k$
- Naïve Bayes → assume **independence:** $P(X \mid Y = k) = \prod_{i=1}^{n} P(X_i \mid Y = k)$

Naïve Bayes

- One version of naïve Bayes with continuous data: assume

$$P(Y = k) = \pi_k \qquad \text{and} \qquad X \mid Y = k \sim N(\mu_k, \sigma^2 I)$$

- Fit the model via MLE (using the training data): under the model above,

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i = k\}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} 1\{Y_i = k\} X_i$$

$$\hat{\sigma}^2 = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \bar{X}_{\cdot j})^2$$

- Beyond the normal model, what does this model assume?
  - Within each class, features have same variance and are **independent**!!
    - Geometrically, this is assuming that the classes are spherically distributed

Linear Discriminant Analysis (LDA)

- In the Gaussian case, let's relax this independence assumption and instead assume

$$X \mid Y = k \sim N(\mu_k, \Sigma_w)$$

  where Σ_w denotes the within-class covariance matrix

- Can again fit model via MLE:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i = k\}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} 1\{Y_i = k\} X_i$$

$$\hat{\boldsymbol{\Sigma}}_w = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:Y_i=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top$$

- Turns out the Bayes classifier under this new assumption is equivalent to LDA
- Can show that the LDA decision boundary is linear in X
  - Consequently, works well when classes are linearly separable

Linear Discriminant Analysis (LDA)

- Another (equivalent) way to think about LDA: decomposition of variance

$$\hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Sigma}}_b + \hat{\boldsymbol{\Sigma}}_w$$

Total variation | Between-class variation | Within-class variation

where

$$\hat{\boldsymbol{\Sigma}}_t = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$$

$$\hat{\boldsymbol{\Sigma}}_b = \frac{1}{n-1}\sum_{k=1}^{K} n_k(\hat{\mu}_k - \bar{X})(\hat{\mu}_k - \bar{X})^\top$$

$$\hat{\boldsymbol{\Sigma}}_w = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:Y_i=k}(X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top$$

Linear Discriminant Analysis (LDA)

- Another (equivalent) way to think about LDA: decomposition of variance

$$\hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Sigma}}_b + \hat{\boldsymbol{\Sigma}}_w$$

Total variation | Between-class variation | Within-class variation

- Beginning of the proof:

$$\hat{\boldsymbol{\Sigma}}_t = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top = \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i:Y_i=k}(X_i - \bar{X})(X_i - \bar{X})^\top$$

$$= \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i:Y_i=k}[(X_i - \hat{\mu}_k) + (\hat{\mu}_k - \bar{X})][(X_i - \hat{\mu}_k) + (\hat{\mu}_k - \bar{X})]^\top$$

Linear Discriminant Analysis (LDA)

- Another (equivalent) way to think about LDA: decomposition of variance

$$\hat{\boldsymbol{\Sigma}}_t \quad = \quad \hat{\boldsymbol{\Sigma}}_b \quad + \quad \hat{\boldsymbol{\Sigma}}_w$$

Total variation | Between-class variation | Within-class variation

- LDA finds a linear projection of the data that maximizes the between-class variation while controlling for the within class variation

$$\max_{v_k} \; v_k^\top \hat{\boldsymbol{\Sigma}}_b v_k \qquad \text{subject to } \; v_k^\top \hat{\boldsymbol{\Sigma}}_w v_k = 1, \quad v_k^\top \hat{\boldsymbol{\Sigma}}_w v_j = 0 \;\; (\forall\, j < k)$$

- This is a *generalized eigenvalue problem*: solution is the eigendecomposition of $\hat{\boldsymbol{\Sigma}}_w^{-1}\hat{\boldsymbol{\Sigma}}_b$

- Why do we care? Enables easy visualization
  - If we put the discriminant directions into a matrix $V = [v_1, \ldots, v_K]$, then the discriminant components XV are the lower-dimensional projections of data that best separate the classes!

Linear Discriminant Analysis (LDA)

- Assumptions of LDA:
  - Implicit multivariate normal assumption: $X \mid Y = k \sim N(\mu_k, \Sigma_w)$
  - Decision boundaries are linear
  - Assumes Σ_w is the same for each class
- We can allow the within class covariance to be different for each class, that is,
$$X \mid Y = k \sim N(\mu_k, \Sigma_k)$$
- This results in quadratic decision boundaries and hence called **quadratic discriminant analysis (QDA)**
- QDA is more flexible than LDA, but requires estimating more parameters
- For both LDA and QDA, if n < p, then can't get Σ_w^{-1}; in this case, add regularization (**regularized discriminant analysis (RDA)**)

Review: Classification methods thus far

	Logistic	Naïve Bayes	LDA	QDA
Pros	• Can do inference (with all the caveats)	• Can choose any likelihood model	• Convenient visualizations • Linearly separable	• Quadratic decision boundaries
Cons	• Problems when p>n (a solution: regularized logistic regression) • Model misspecification?	• Assumes that features are independent (a very strong assumption) • Model misspecification?	• Problems when p>n (a solution: RDA) • Model misspecification? Non-normal or non-linear decision boundaries?	• Problems when p>n (a solution: RDA) • Requires larger n to estimate more parameters adequately (compared to LDA) • Model misspecification? Non-normal or non-linear decision boundaries?

- LDA is more *efficient* than logistic regression if X is Gaussian (i.e., LDA requires fewer samples to do well) whereas logistic regression is better than LDA if X is not Gaussian

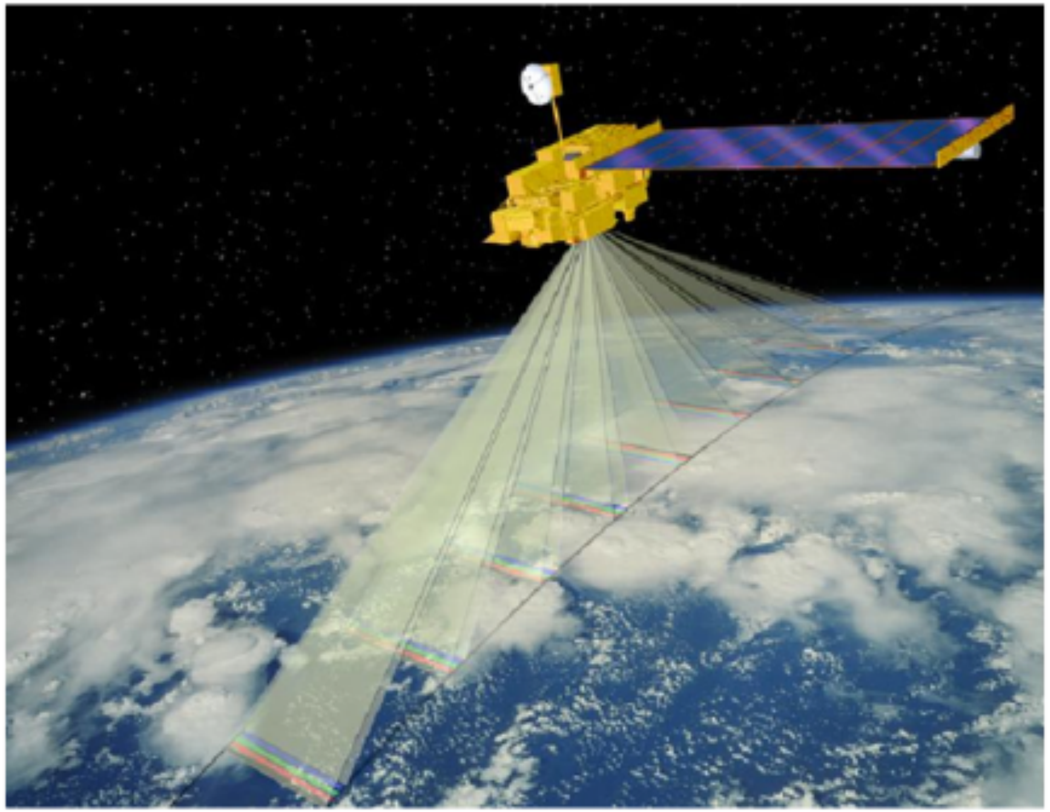
K Nearest Neighbors (KNN) Classifier

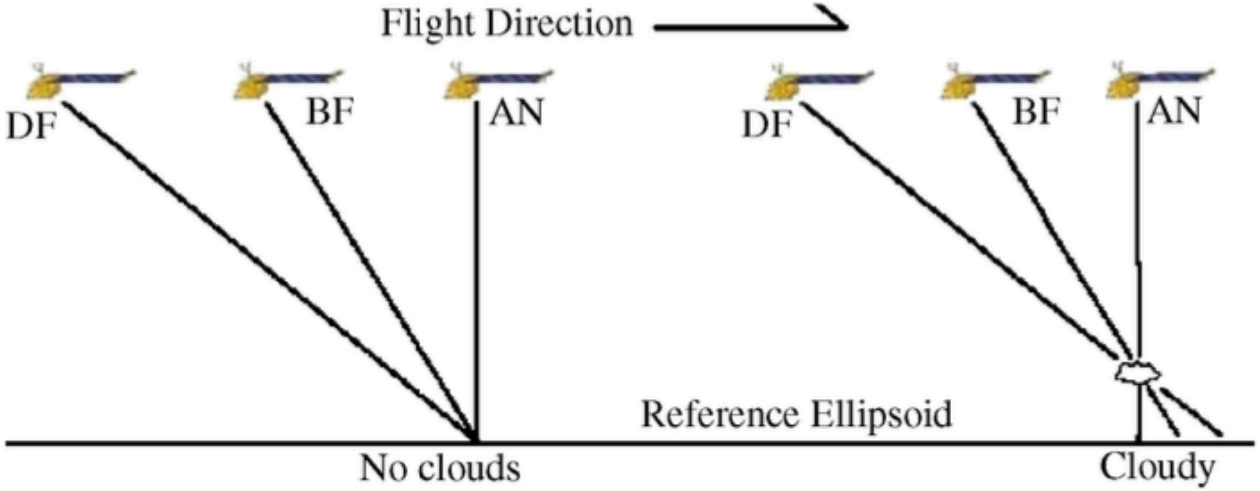
- Let's take a purely algorithmic approach to classification
  - Maybe this is good or bad?
- For each test sample (or for each sample you want to make a prediction on):
  - Find the K "closest" neighbors
    - How do we define closest? Need to choose d(x, y)
  - Take majority vote from K closest neighbors
- Advantages: flexible, data-adaptive, simple, easy
- Disadvantages: curse of dimensionality
- In R: class::knn()

Next time

- SVMs
- Random Forests
- Ensembles
- Evaluation?

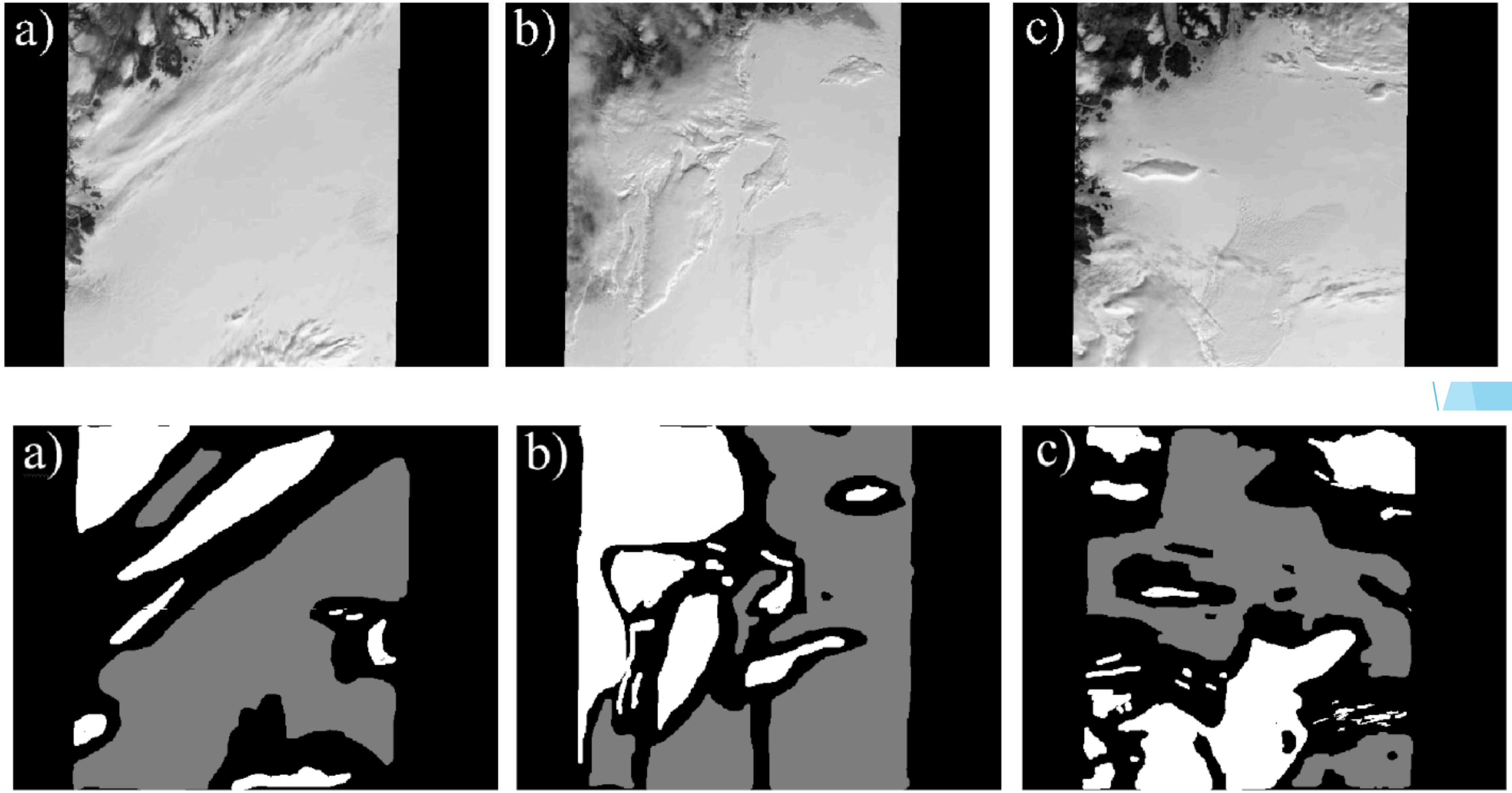
Lab 4: Remote Sensing/Cloud Data





Feature engineering:
CORR, SD, NDAI

Lab 4: Remote Sensing/Cloud Data



Lab 4: Things to think about carefully

- Which methods/models? Are they well-suited for this data? Why or why not? What are the advantages/disadvantages and assumptions of the method(s) that you chose?
  - This can help you better identify the limitations of your prediction algorithm
- Data splitting scheme? This is very important for generalizability
- Post-hoc EDA? Can provide insights into how to improve your prediction

Lab 4 Groups

1	Cam Adams	Malvika Rajeev	Sohum Datta
2	Chao Zhang	Facu Sapienza	Jiaxi Liu
3	Corrine Elliott	Sam Stein	Yihuan Song
4	Katherine Kempfert	Phil Ryjanovsky	
5	Partow Imani	Yanting Pan	Yiyi He
6	Kanaad Deodhar	Liang Zhang	Mike Janson
7	Chenxing Wu	Ella Hiesmayr	Namita Trikannad
8	Dodo Qian	Robbie Netzorg	Teng Li
9	Aya Amanmyradova	Brooke Staveland	Shubei Wang
10	Spencer Wilson	Ziyang Zhou	