STAT 215A Fall 2023 Week 13

Chengzhong Ye

Announcements

- How’s final project going?

Outline for today

- Classification algorithms
  - Logistic regression
  - Naive Bayes
  - Discriminant analysis
  - Evaluation metrics

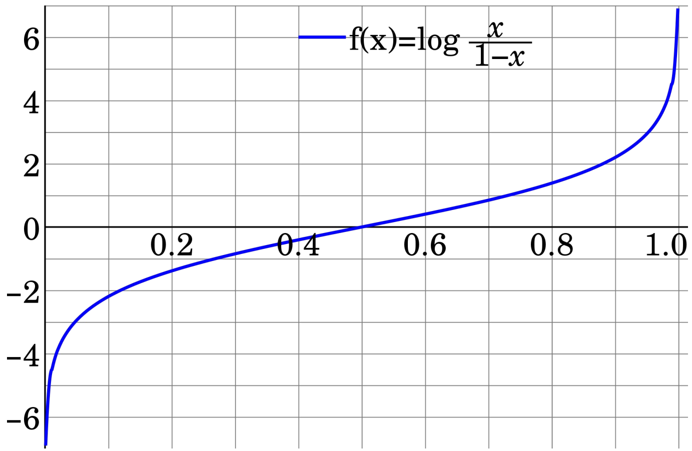
Why classification and not regression?

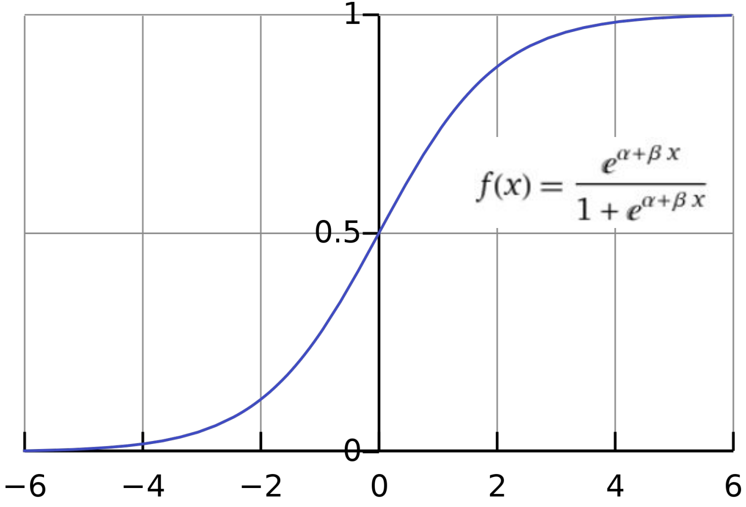
- Suppose we have data $X_1, \ldots, X_n$ and categorical responses $y_1, \cdots, y_n$, i.e. $y_i \in 1, \ldots, K$.
- Problems with regression:
  - Hard to assign numeric values to categories
  - Usually no ordering of the categories
  - Even if categories are ordered, not necessarily equally spaced

Logistic regression

Assume there are two classes and $y_i | x_i \sim \text{Bernoulli}(\pi_i)$ are independent with

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i \iff \pi_i = \frac{\exp\{\alpha + \beta x_i\}}{1 + \exp\{\alpha + \beta x_i\}}$$





Find MLE via Newton-Raphson / IRLS. **`glmnet`** can fit large logistic regression models efficiently.

Logistic Regression Extensions

- What if more than 2 classes?
  - Multinomial logistic regression
- What if $p > n$ (or p large)?
  - Regularized logistic regression: $\max_{\alpha,\beta} \ell(\alpha, \beta, X) - \lambda \underbrace{q(\beta)}_{\text{Penalty, e.g. } L^1,\, L^2}$
- What assumptions are you making?
  - Linear relationship between covariates and log-odds.
  - Correlated predictors can inflate variance and bias of coefficients

Modeling via class conditional densities

$\in \mathbb{R}^p$

If we know the class posterior distribution $P(Y = k|X)$, then we could just predict the class k with the highest probability given the observation.

- Say $f_k(x)$ is the conditional density of an observation within the class k
- Call π_k the prior probability of the class k and assume $\sum_{k=1}^{K} \pi_k = 1$

Modeling via class conditional densities

$\in \mathbb{R}^p$

If we know the class posterior distribution $P(Y = k|X)$, then we could just predict the class k with the highest probability given the observation.

- Say $f_k(x)$ is the conditional density of an observation within the class k
- Call π_k the prior probability of the class k and assume $\sum_{k=1}^{K} \pi_k = 1$

Then, using Bayes rule we have $P(Y = k|X) = \dfrac{f_k(x)\pi_k}{\sum_l f_l(x)\pi_l}$

Naive Bayes

Assumes that given the class label, the features are independent!

$$f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$$

- E.g., model the covariates via independent Gaussians: $X|Y = k \sim N(\mu_k, \sigma^2 \mathrm{I})$
- This makes estimation much simpler, and can actually work well in practice in spite of this strong assumption.

- Maximum a priori estimator

Linear discriminant analysis (LDA)

LDA is based upon modeling the class conditional density $f_k(x)$ via a Gaussian with **equal variance** within each class (but not necessarily independent).

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

within class covariance matrix, common across classes

Linear discriminant analysis (LDA)

LDA is based upon modeling the class conditional density $f_k(x)$ via a Gaussian with **equal variance** within each class (but not necessarily independent).

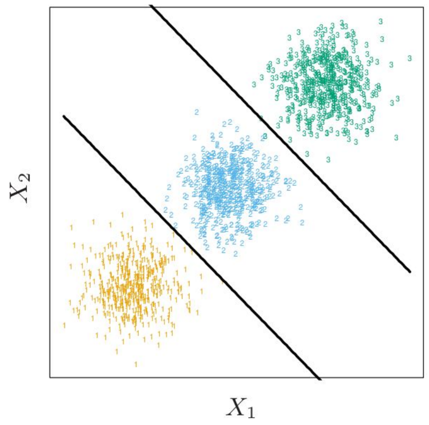
$$X|Y = k \sim N(\mu_k, \Sigma_w)$$

within class covariance matrix, common across classes

- **Exercise**: show that, for this model, we have

$$\log \frac{P(Y=k|X)}{P(Y=l|X)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k - \mu_l)^\top \Sigma^{-1}(\mu_k + \mu_l) + x^\top \Sigma^{-1}(\mu_k - \mu_l)$$

linear in x!



Linear discriminant analysis (LDA)

LDA is based upon modeling the class conditional density $f_k(x)$ via a Gaussian with **equal variance** within each class (but not necessarily independent).

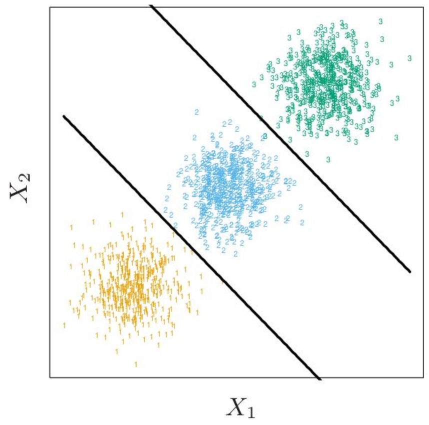
$$X|Y = k \sim N(\mu_k, \Sigma_w)$$

within class covariance matrix, common across classes

- **Exercise**: show that, for this model, we have

$$\log \frac{P(Y=k|X)}{P(Y=k|X)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k - \mu_l)^\top \Sigma^{-1}(\mu_k + \mu_l) + x^\top \Sigma^{-1}(\mu_k - \mu_l)$$

linear in x!



- We can fit the parameters via MLE:

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

LDA as decomposition of variance

Can think about LDA as a decomposition of variance:

$$\underset{\text{Total variation}}{\hat{\mathbf{\Sigma}}_t} = \underset{\text{Between-class variation}}{\hat{\mathbf{\Sigma}}_b} + \underset{\text{Within-class variation}}{\hat{\mathbf{\Sigma}}_w}$$

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

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

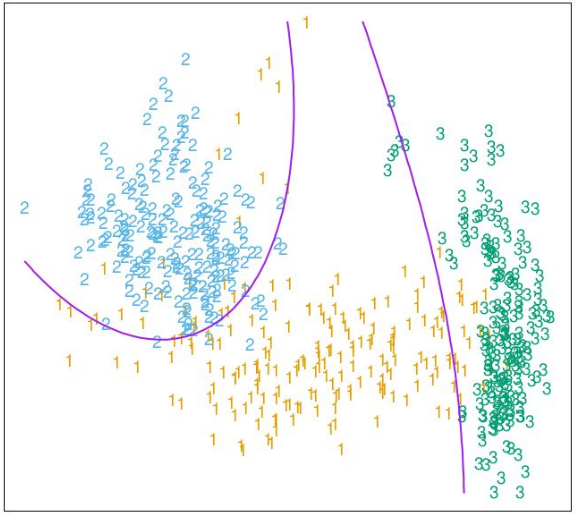
- Collect into a matrix $V = [v_1, \ldots, v_K]$ and look at discriminant components XV
  - Low-dim projection of data that best separates the classes!

LDA vs. Logistic Regression (LR)

The two methods seem to be very similar, but get to their results by very different methods, with important implications.

- Assumptions:
  - **LR makes fewer assumptions** and is therefore more general.
  - The additional assumptions imposed by **LDA leads to lower variance** of estimates (especially when true data is Gaussian).
- Robustness
  - Assumptions make **LDA more sensitive to outliers**
  - **LR downweights outliers** far from the decision boundary, making it more robust
- In practice, results are very similar, but LR may be a safer bet

Quadratic discriminant analysis (QDA)



- When classes cannot be separated by a hyperplane, one option is to use LDA with quadratic features.
- Another is to relax the equal variance-covariance constraint, which results in QDA:

$$X|Y = k \sim N(\mu_k, \Sigma_k)$$

- Now we have to estimate separate covariance matrices for each class which can result in many more parameters.
- Another variant: Regularized Discriminant Analysis
  - Shrink the separate covariance matrices toward a common one

Summary so far

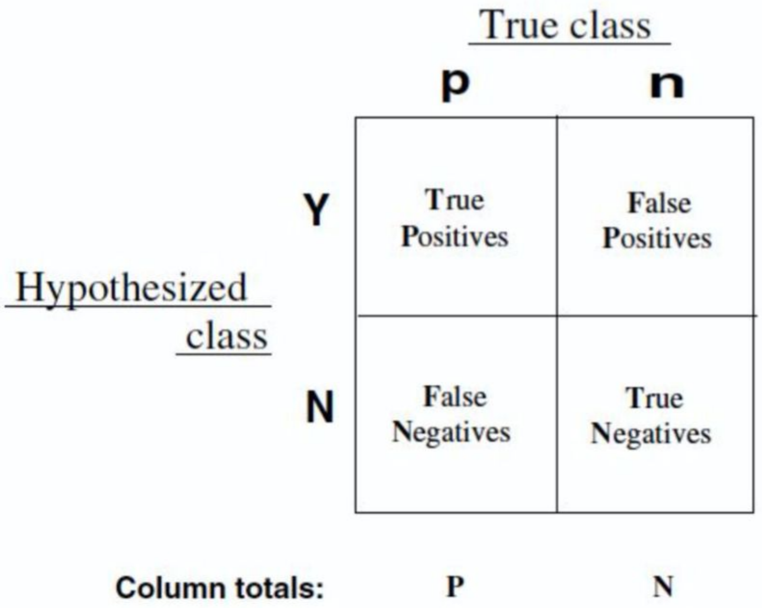
	Logistic	Naïve Bayes	LDA	QDA
Pros	• Can do inference (with all the caveats)	• Can choose any likelihood model	• Convenient visualizations • Linearly separable	• Quadratic decision boundaries
Cons	• Problems when p>n (a solution: regularized logistic regression) • Model misspecification?	• Assumes that 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?

Evaluation metrics for classification

How to evaluate your classification methods?

- Going beyond classification error
- What if we have class imbalance?
    - For example, if we take a sample of 100 people and only 10 have the disease, then always predicting healthy gives 90% classification accuracy!
    - We can do better.

Confusion matrix



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

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

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

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

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

Source: Fawcett (2005)

Confusion matrix

		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

True class

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

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

Precision-recall curve

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

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

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

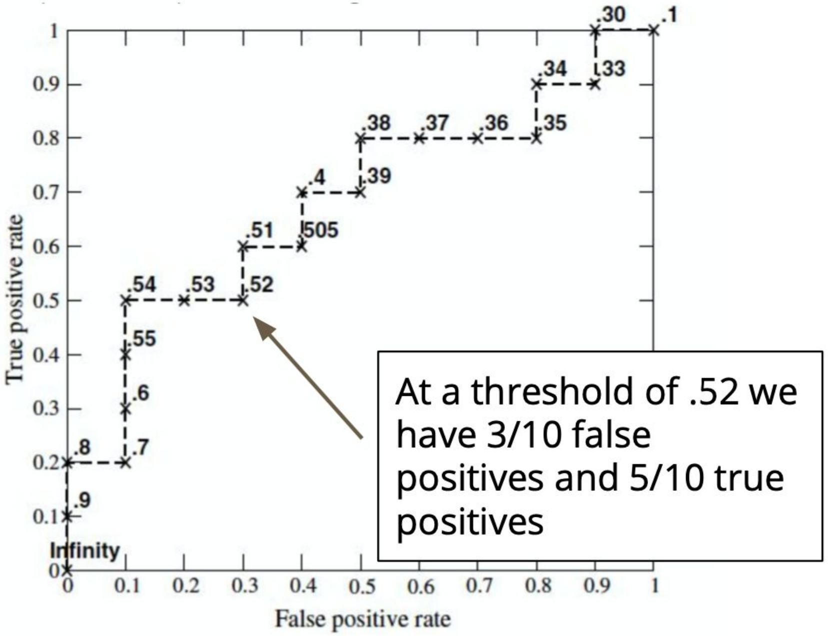
Fig. 1. Confusion matrix and common performance metrics calculated from it.

Source: Fawcett (2005)

Receiver operating characteristics (ROC) curve

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)

Area under the curve

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)

AUC in practice

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)

ROC vs Precision-Recall (PR) Curves

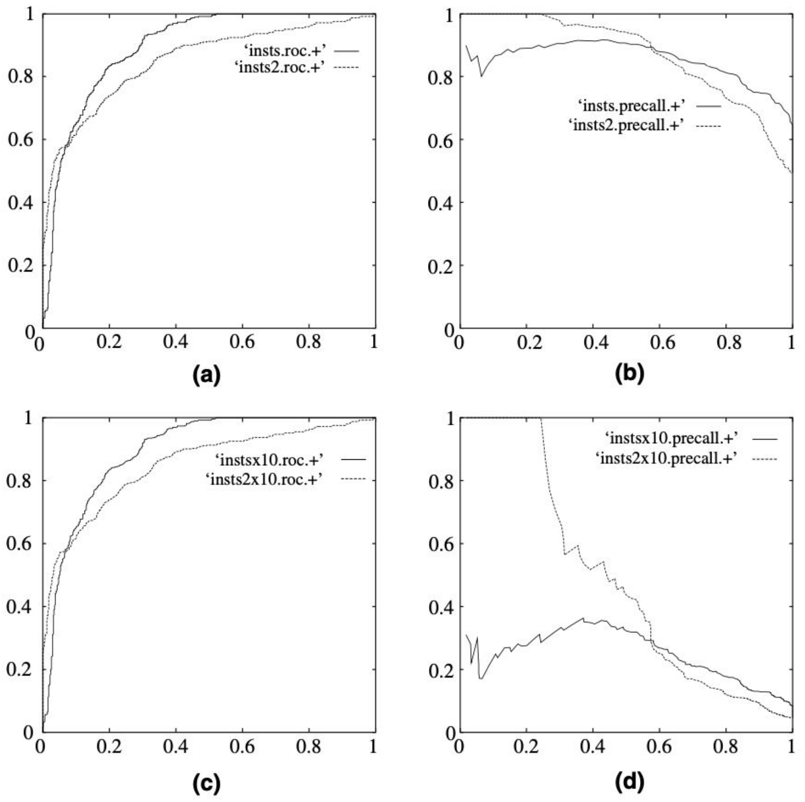


Fig. 5. ROC and precision-recall curves under class skew. (a) ROC curves, 1:1; (b) precision-recall curves, 1:1; (c) ROC curves, 1:10 and (d) precision-recall curves, 1:10.

ROC vs PR curves

- Generally, precision-recall curves are preferred when there is class imbalance
- ROC curves tend to paint an overly optimistic view of the model on datasets with class imbalance
- PR calculations do not involve the true negatives rate and hence do not typically present such an optimistic view