

Lecture 11: LDA and Logistic regression

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 1/23

Announcements

- Homework 3 due next Tuesday

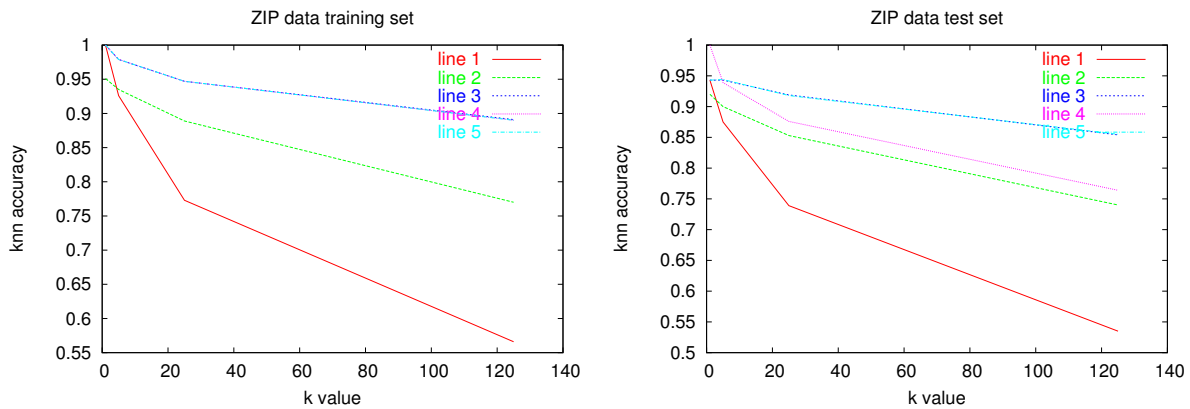
CSI 5v93: Introduction to machine learning, Lecture 11 – p. 2/23

Questions?

Comments on homework 2

- Be very clear in your descriptions.
- Use lots of graphics to explain your data and support your conclusions.
- Explain the graphics, too!
- Make your document pretty. Tables of numbers should be well-formatted. Use the `verbatim` and `tabular` environments.
- Evaluate your answers.
- Ask questions.
- Start early.

k -nearest neighbor results from 5 students



Comments:

- The k -nn algorithm is deterministic and well-defined.
- 1-nearest neighbor should get 100% accuracy on training data.
- If you get 100% accuracy on test data, there might be a problem.

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 5/23

Making clear, supportable statements

Which is more clear? Which is supported?

- “ k -nn is more efficient to learn.”
- “ k -nn is efficient for learning, but slow for classification, compared to linear regression. k -nn requires no *training* time, while linear regression requires computation of the beta values from data. However, for classifying new data, linear regression is a simple multiplication, but k -nn must search the entire dataset and perform a sort.”
- “ k -nn has higher accuracy.”
- “ k -nn achieved 94% accuracy on unseen test data for $k = 5$. This accuracy is for classifying on examples of all digits (0-9). Linear regression achieved 92% classification accuracy on unseen test data for classifying between digits 3 and 8.”

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 6/23

Chapter 4: Linear methods for classification

- 4.1 – Introduction
- 4.2 – Linear regression of an indicator matrix
- 4.3 – Linear discriminant analysis
- 4.4 – Logistic regression
- 4.5 – Separating hyperplanes

Linear decision boundaries for classification (4.1)

Classifier on real-valued data

Decision boundaries

Linear decision boundaries

Linear regression for classification

- assume that \mathbf{Y} is a $n \times k$ indicator matrix
- learn k linear models, one for each class
- problem: masking

Discriminant functions and linear decision boundaries

Recall our problem framework for classification: k discriminant functions $\delta_k(x)$

The decision boundary between classes a and b :

$$\{x \mid \delta_a(x) = \delta_b(x)\}$$

We are interested in decision boundaries which are linear.

The discriminant function could be linear to achieve this, but it is not necessary.

Posterior probabilities

We want to classify to the class that gives the largest posterior probability:

For class l :

- Prior probability: π_l
- Model probability: $f_l(x)$

Using Bayes' rule, the class posterior probability is:

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}$$

Linear discriminant analysis

Log-odds:

$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = l|X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

The boundary is where the log-odds equals 0.

LDA and Gaussian probabilities

Gaussian probability distribution in one dimension $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Spherical Gaussian in d dimensions $N(\mu, \sigma^2 \mathbf{I})$:

$$f(x) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp \left\{ -\frac{\|x - \mu\|^2}{2\sigma^2} \right\}$$

General Gaussian in d dimensions $N(\mu, \Sigma)$:

$$f(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 13/23

LDA and Gaussian probabilities

Assume that $\Sigma_k = \Sigma$ for all classes, which simplifies the log-odds:

$$\begin{aligned} \log \frac{\Pr(G = k | X = x)}{\Pr(G = l | X = x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\ &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) \end{aligned}$$

The boundary is (still) where the log-odds equals 0.

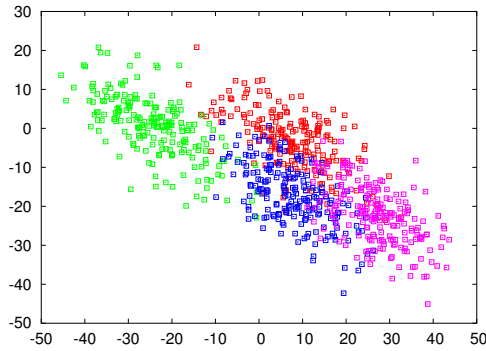
The boundary has a linear dependence on x .

From this we can see that the *linear discriminant function* is:

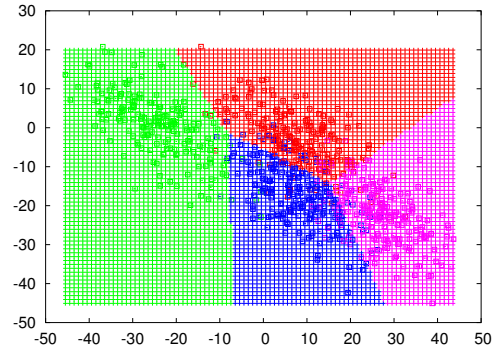
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 14/23

Examples of LDA (in Matlab)



Data from 4 classes (colored red, green, blue, purple), each distributed as 2-d Gaussian with same covariance



After learning LDA parameters – the classification regions

Note that the decision boundaries are linear.

Learning (estimating) the parameters

We do not know π_k , μ_k , or Σ .

We must learn them from the data:

$$\begin{aligned}\hat{\pi}_k &= \frac{N_k}{N} \\ \hat{\mu}_k &= \frac{1}{N_k} \sum_{g_i=k} x_i \\ \hat{\Sigma} &= \frac{1}{N - k} \sum_{k=1}^K \sum_{g_i=k} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)\end{aligned}$$

Related to LDA

QDA (quadratic discriminant analysis):

- covariances (Σ_k) are not restricted, and are estimated for each class
- decision boundaries are not linear
- discriminant functions $\delta_k(x)$ are quadratic (not linear)
- we can use LDA with extra quadratic features to produce results similar to QDA

Regularized discriminant analysis: cross between LDA and QDA

$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma}$$

Logistic regression (4.4)

Logistic regression is similar to LDA in concept. They both:

- classify to the largest posterior probability $\Pr(G = k|X = x)$
- form linear decision boundaries between classes
- can use log-odds to express class boundaries; for example:

$$\log \frac{\Pr(G = a|X = x)}{\Pr(G = b|X = x)} = \alpha_0 + \alpha^T x$$

However, they are different in several ways.

Logistic regression model

Logistic regression models the log-odds directly with respect to the last (K th) class:

$$\begin{aligned}\log \frac{\Pr(G = 1|X = x)}{\Pr(G = K|X = x)} &= \beta_{10} + \beta_1^T x \\ \log \frac{\Pr(G = 2|X = x)}{\Pr(G = K|X = x)} &= \beta_{20} + \beta_2^T x \\ &\vdots \\ \log \frac{\Pr(G = K - 1|X = x)}{\Pr(G = K|X = x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x\end{aligned}$$

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 19/23

Logistic regression model

From the log-odds for each class (with respect to the last class):

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = K|X = x)} = \beta_{10} + \beta_1^T x$$

We can derive the posterior class probabilities for each class:

$$\begin{aligned}\Pr(G = k|X = x) &= \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^T x)} \quad (\text{for } k = 1 \dots K - 1) \\ \Pr(G = K|X = x) &= \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + \beta_i^T x)}\end{aligned}$$

Note the probabilistic nature of these functions:

- $\Pr(G|X)$ is between 0 and 1
- $\sum_{k=1}^K \Pr(G = k|X = x) = 1$

CSI 5v93: Introduction to machine learning, Lecture 11 – p. 20/23

Fitting Logistic regression models

What are the best parameters (the β values that fit the data the best)?

Formulate as a maximum-likelihood problem:

- form the likelihood function (product of probabilities of all training points according to the model)
- take the derivatives of the likelihood function with respect to the parameters (the β s)
- use the derivatives to find the β s that maximize the likelihood

Sometimes maximum likelihood gives a simple closed-form solution; sometimes not.

With logistic regression, the answers are not closed-form, and we must use numerical methods to find the best parameter values.

LDA/Logistic regression differences

	LDA	Logistic Regression
model assumed	Gaussian for each class, w/common covariance	none
models $\Pr(X)$?	yes; mixture of Gaussians	no
learning params	closed form (simple)	max-likelihood using numerical optimization (harder)
probs in $[0,1]$?	no	yes

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification