DS 5220

Supervised Machine Learning and Learning Theory

Alina Oprea
Associate Professor, CCIS
Northeastern University

September 23 2019

Logistics

- HW 1 is due today, Sept. 23
- Midterm
  - Monday, Oct. 28
- Final exam
  - In the last class, Wed. Dec. 4
- Project report due during the final exam week
- Project milestone
  - Form teams of 2-3 people
  - Submit project proposal on Oct. 16
  - Project pitch on Oct. 21 in class

Outline

- Linear classifier
  - LDA on one dimension
  - Linear discriminant functions
  - Multi-variate LDA
- Bias-variance tradeoff
  - Derivation for linear regression

LDA

- Classify to one of k classes (multi-class)
- LDA uses Bayes Theorem to estimate it
  - $\mathrm{P}[Y = k|X = x] = \frac{\mathrm{P}[X = x|Y = k]\mathrm{P}[Y=k]}{\mathrm{P}[X=x]}$
  - Let $\pi_k = \mathrm{P}[Y = k]$ be the prior probability of class k and $f_k(x) = \mathrm{P}[X = x|Y = k]$
- Generative model
  - Given X and Y, learns the joint probability $P(X, Y)$
- Discriminative model
  - Given X and Y, learns a decision function for classification

LDA

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

Assume $f_k(x)$ is Gaussian!
Unidimensional case (d=1)

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}.$$

Assumption: $\sigma_1 = \ldots \sigma_k = \sigma$

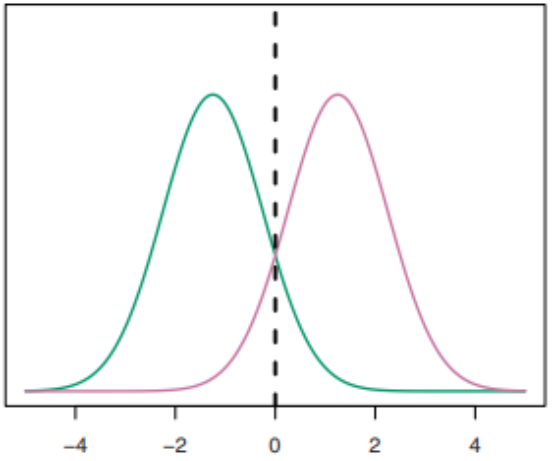
LDA decision boundary

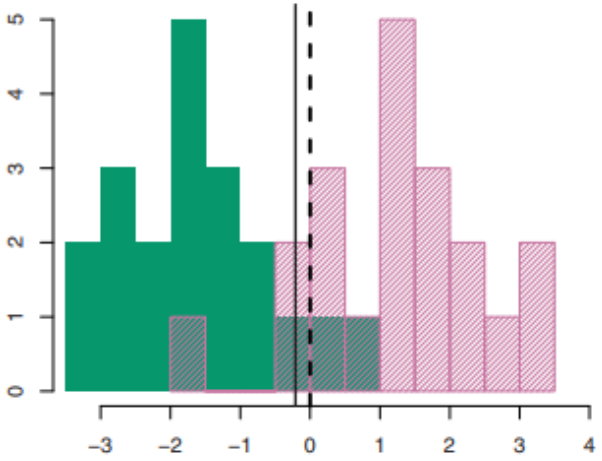
Pick class k to maximize

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Example: $k = 2, \pi_1 = \pi_2$

Classify as class 1 if $x > \frac{\mu_1 + \mu_2}{2}$





True decision boundary

Estimated decision boundary

LDA in practice

Given training data $(x_i, y_i), i = 1, \ldots, n, y_i \in \{1, \ldots, K\}$

1. Estimate mean and variance

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

2. Estimate prior

$$\hat{\pi}_k = n_k/n.$$

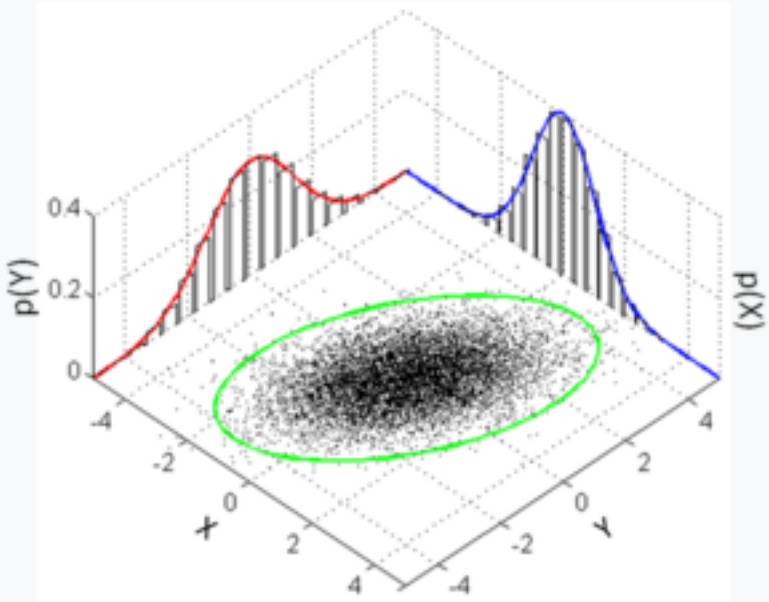
Given testing point x, predict k that maximizes:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

Multi-Variate Normal

Multivariate normal

Probability density function



Many sample points from a multivariate normal distribution with $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix}$, shown along with the 3-sigma ellipse, the two marginal distributions, and the two 1-d histograms.

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with *k*-dimensional mean vector

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]^{\mathrm{T}},$$

and $k \times k$ covariance matrix

$$\Sigma_{i,j} =: \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$$

$$\boldsymbol{\Sigma} =: \mathrm{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}] = [\mathrm{Cov}[X_i, X_j]; 1 \leq i, j \leq k].$$

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k)}.$$

Multi-variate LDA

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

Assume $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$

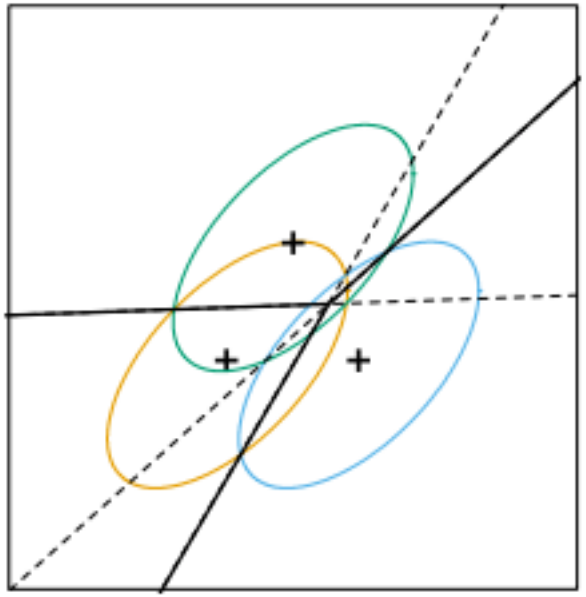
$$\begin{aligned}\log \frac{\Pr(Y = k|X = x)}{\Pr(Y = l|X = x)} &= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} \\ &= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \boldsymbol{\Sigma}^{-1}(\mu_k - \mu_\ell) \\ &\quad + x^T \boldsymbol{\Sigma}^{-1}(\mu_k - \mu_\ell),\end{aligned}$$

Linear decision boundary between classes *k* and *l*

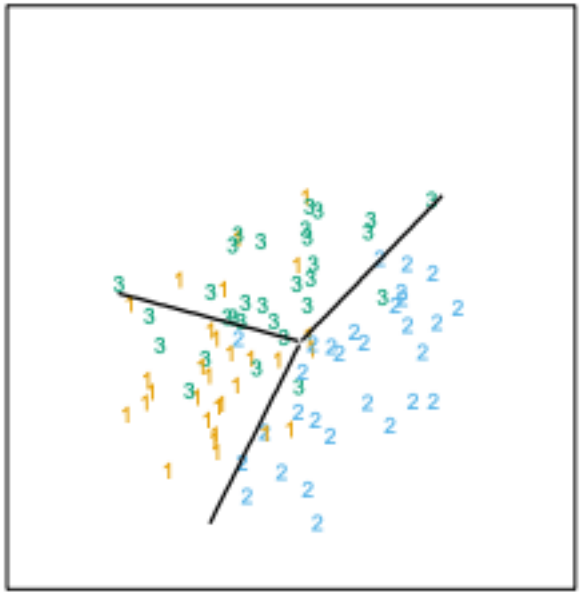
Linear discriminant functions $\delta_k(x) = x^T \boldsymbol{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k + \log \pi_k$

Given x, classify to class k: $argmax_k \delta_k(x)$

Example 3 classes



3 Normal distributions with same co-variance, but different means



LDA decision boundary

Multi-variate LDA

Given training data $(x_i, y_i), i = 1, \ldots, n, y_i \in \{1, \ldots, K\}$

1. Estimate mean and variance

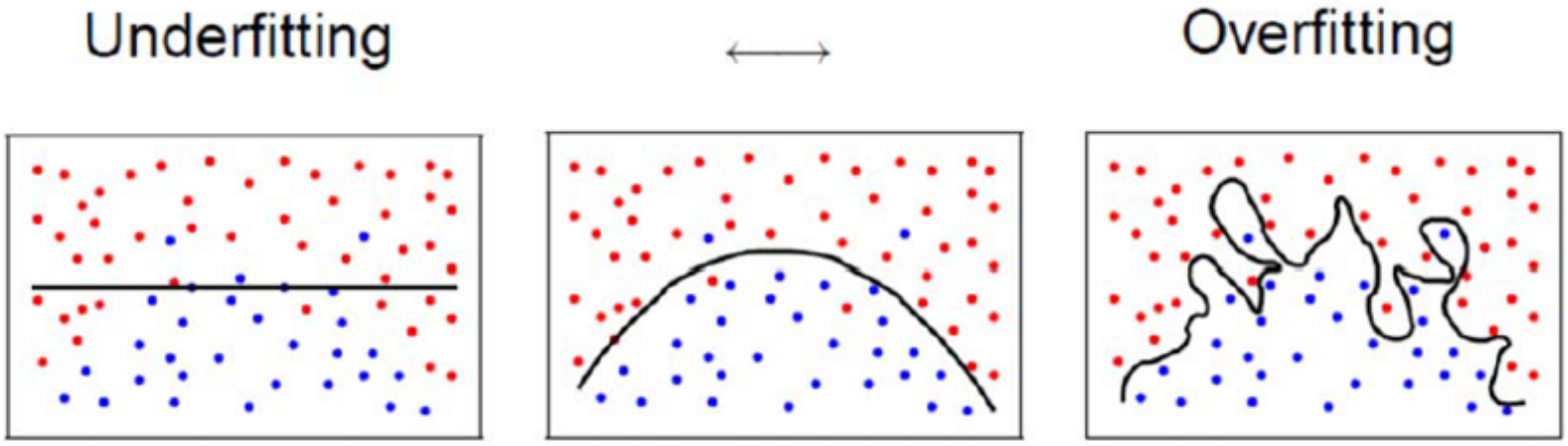
- $\hat{\pi}_k = N_k/N$, where N_k is the number of class-k observations;
- $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$;
- $\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N - K)$.

2. Estimate prior

Given testing point x, predict k that maximizes:

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

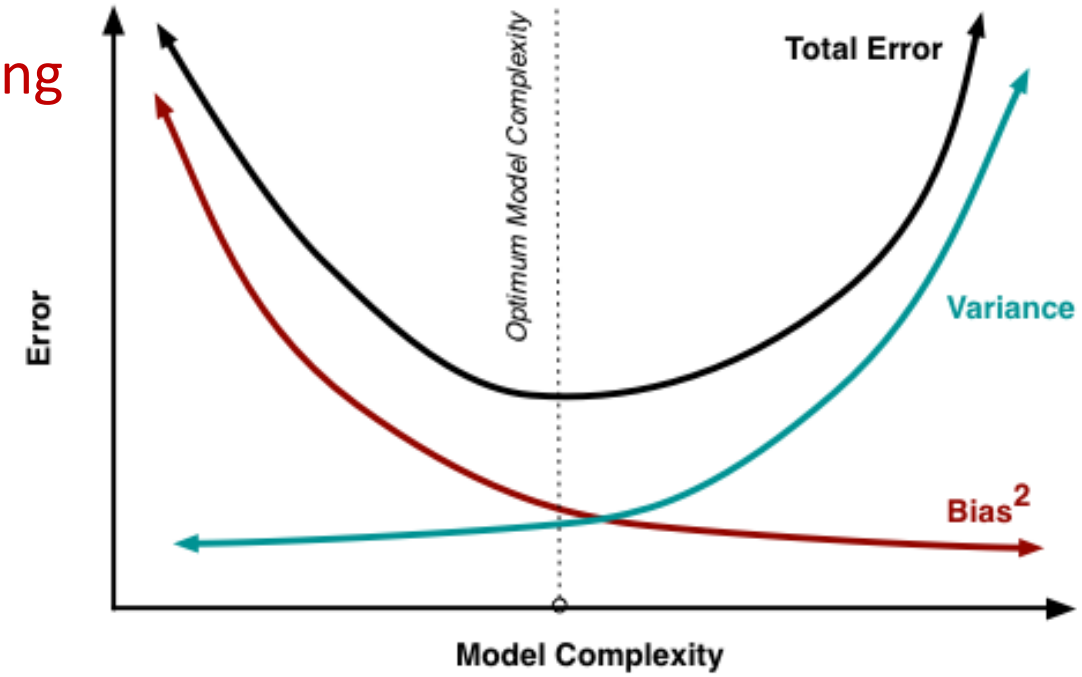
Generalization in ML



- Goal is to generalize well on new testing data
- Risk of overfitting to training data
  - MSE close to 0, but performs poorly on test data

Bias-Variance Tradeoff

Under-fitting



Over-fitting

- Bias = Difference between estimated and true models
- Variance = Model difference on different training sets

MSE is proportional to Bias + Variance

Bias-Variance Decomposition

- Assume that $y = f(\boldsymbol{x}) + \epsilon$
  - Noise ϵ is sampled from a normal distribution with 0 mean and variance σ^2: $\epsilon \sim N(0, \sigma^2)$
  - Noise lower-bounds the performance we can achieve

- Recall the following objective function:

$$J(\theta) = \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2$$

- We can re-write this as the expected value of the squared error: $\mathrm{E}\left(y - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right)^2$

$f(x)$: True model (not necessarily linear)
$h_\theta(x)$: Learned model (linear)

Bias-Variance Decomposition

$$\begin{aligned}
\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] &= \mathrm{E}[(y - f(\boldsymbol{x}) + f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{E}[(y - f(\boldsymbol{x}))^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\quad + 2\,\mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))(y - f(\boldsymbol{x}))] \\
&= \mathrm{E}[(y - f(\boldsymbol{x}))^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\quad + 2\,(\mathrm{E}[f(\boldsymbol{x})h_{\boldsymbol{\theta}}(\boldsymbol{x})] + \mathrm{E}[yf(\boldsymbol{x})] - \mathrm{E}[yh_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathrm{E}[f(\boldsymbol{x})^2])
\end{aligned}$$

cancels cancels

Therefore,

$$\begin{aligned}
\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] &= \mathrm{E}[(y - f(\boldsymbol{x}))^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{E}[\epsilon^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]
\end{aligned}$$

Aside:
Definition of Variance
$\mathrm{var}(z) = \mathrm{E}[(z - \mathrm{E}[z])^2]$

This is actually $\mathrm{var}(\epsilon)$, since mean is 0

7

Bias-Variance Decomposition

$$
\begin{aligned}
\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] &= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] + \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2] + \mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\quad + 2\mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])] \\
&= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2] + \mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\quad + 2\left(\mathrm{E}[f(\boldsymbol{x})\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]] - \mathrm{E}[\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]^2] - \mathrm{E}[f(\boldsymbol{x})h_{\boldsymbol{\theta}}(\boldsymbol{x})] + \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]]\right)
\end{aligned}
$$

cancels cancels

Therefore,

$$\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] = \underbrace{\mathrm{var}(\epsilon)}_{\text{noise}} + \underbrace{\mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2]}_{\text{bias}} + \underbrace{\mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]}_{\text{variance}}$$

$$\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] = \mathrm{bias}(h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2 + \mathrm{var}(h_{\boldsymbol{\theta}}(\boldsymbol{x})) + \sigma^2$$

Review

- LDA
  - Example of generative model
  - Use Bayes Theorem to estimate the probability that label is from each class k
  - Assumes normal distribution of features in each class
- Bias-Variance tradeoff for linear regression
  - Decompose expectation of testing error as a sum of squared bias, variance, and noise term
  - Shows that bias and variance need to be simultaneously minimized

Acknowledgements

- Slides made using resources from:
  - Andrew Ng
  - Eric Eaton
  - David Sontag
- Thanks!