Gaussian Discriminant Analysis

Isaac Osei Nyantakyi, PH.D

Generative vs Discriminative (Recap)

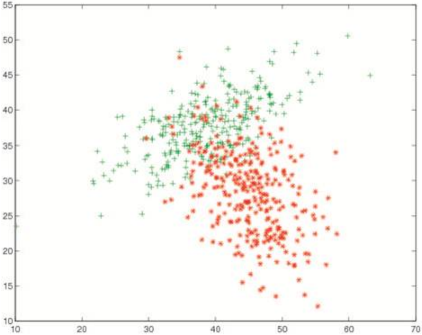
Two approaches to classification:

- Discriminative approach: estimate parameters of decision boundary/class separator directly from labeled examples.
  - Model $p(t|\mathbf{x})$ directly (logistic regression models)
  - Learn mappings from inputs to classes (linear/logistic regression, decision trees etc)
  - Tries to solve: How do I separate the classes?
- Generative approach: model the distribution of inputs characteristic of the class (Bayes classifier).
  - Model $p(\mathbf{x}|t)$
  - Apply Bayes Rule to derive $p(t|\mathbf{x})$.
  - Tries to solve: What does each class "look" like?

Classification: Diabetes Example

- Gaussian discriminant analysis (GDA) is a Bayes classifier for continuous-valued inputs.

- Observation per patient: White blood cell count & glucose value.



- $p(\mathbf{x} \mid t = k)$ for each class is shaped like an ellipse

  $\Longrightarrow$ we model each class as a multivariate Gaussian

Gaussian Discriminant Analysis

- Gaussian Discriminant Analysis in its general form assumes that $p(\mathbf{x}|t)$ is distributed according to a multivariate Gaussian distribution
- Multivariate Gaussian distribution:

$$p(\mathbf{x} \,|\, t = k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right]$$

  where $|\boldsymbol{\Sigma}_k|$ denotes the determinant of the matrix.
- Each class k has associated mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$
- How many parameters?
  - Each $\boldsymbol{\mu}_k$ has D parameters, for DK total.
  - Each $\boldsymbol{\Sigma}_k$ has $\mathcal{O}(D^2)$ parameters, for $\mathcal{O}(D^2K)$

GDA: Learning

- Learn the parameters for each class using maximum likelihood
- For simplicity, assume binary classification

$$p(t \mid \phi) = \phi^t (1 - \phi)^{1-t}$$

- You can compute the ML estimates in closed form (ϕ and $\boldsymbol{\mu}_k$ are easy, $\boldsymbol{\Sigma}_k$ is tricky)

$$\begin{aligned}
\phi &= \frac{1}{N} \sum_{i=1}^{N} r_1^{(i)} \\
\boldsymbol{\mu}_k &= \frac{\sum_{i=1}^{N} r_k^{(i)} \cdot \mathbf{x}^{(i)}}{\sum_{i=1}^{N} r_k^{(i)}} \\
\boldsymbol{\Sigma}_k &= \frac{1}{\sum_{i=1}^{N} r_k^{(i)}} \sum_{i=1}^{N} r_k^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^\top \\
r_k^{(i)} &= \mathbb{1}[t^{(i)} = k]
\end{aligned}$$

GDA Decision Boundary

- Recall: for Bayes classifiers, we compute the decision boundary with Bayes' Rule:

$$p(t \,|\, \mathbf{x}) = \frac{p(t)\, p(\mathbf{x} \,|\, t)}{\sum_{t'} p(t')\, p(\mathbf{x} \,|\, t')}$$

- Plug in the Gaussian $p(\mathbf{x} \,|\, t)$:

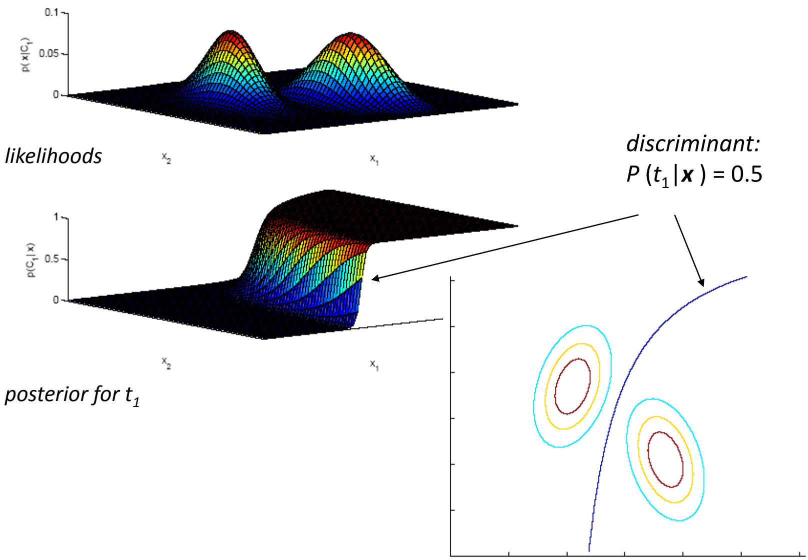
$$\begin{aligned}
\log p(t_k|\mathbf{x}) &= \log p(\mathbf{x}|t_k) + \log p(t_k) - \log p(\mathbf{x}) \\
&= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \\
&\quad + \log p(t_k) - \log p(\mathbf{x})
\end{aligned}$$

- Decision boundary:

$$(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) = (\mathbf{x} - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x} - \boldsymbol{\mu}_\ell) + \text{Const}$$

- What's the shape of the boundary?
  - We have a quadratic function in $\mathbf{x}$, so the decision boundary is a conic section!

GDA Decision Boundary



GDA Decision Boundary

- Our equation for the decision boundary:

$$(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) = (\mathbf{x}-\boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell) + \text{Const}$$

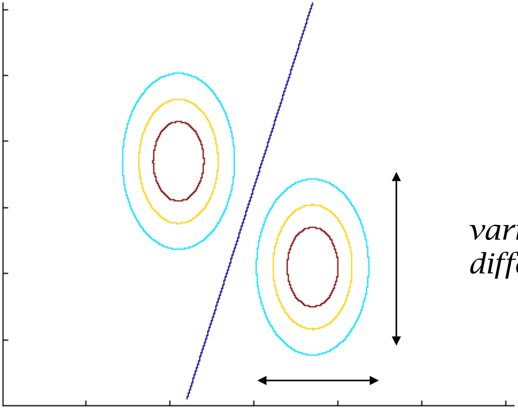
- Expand the product and factor out constants (w.r.t. $\mathbf{x}$):

$$\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x} - 2\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x} = \mathbf{x}^\top \boldsymbol{\Sigma}_\ell^{-1}\mathbf{x} - 2\boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}_\ell^{-1}\mathbf{x} + \text{Const}$$

- What if all classes share the same covariance $\boldsymbol{\Sigma}$?
  - We get a linear decision boundary!

$$\begin{aligned} -2\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} &= -2\boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \text{Const} \\ (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} &= \text{Const} \end{aligned}$$

GDA Decision Boundary: Shared Covariances



variances may be different

GDA vs Logistic Regression

- Binary classification: If you examine $p(t = 1 \mid \mathbf{x})$ under GDA and assume $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, you will find that it looks like this:

$$p(t \mid \mathbf{x}, \phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x} - b)}$$

  where $(\mathbf{w}, b)$ are chosen based on $(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$.

- Same model as logistic regression!

GDA vs Logistic Regression

When should we prefer GDA to logistic regression, and vice versa?

- GDA makes a stronger modeling assumption: assumes class-conditional data is multivariate Gaussian
  - If this is true, GDA is asymptotically efficient (best model in limit of large N)
  - If it's not true, the quality of the predictions might suffer.
- Many class-conditional distributions lead to logistic classifier.
  - When these distributions are non-Gaussian (i.e., almost always), LR usually beats GDA
- GDA can handle easily missing features (how do you do that with LR?)

Gaussian Naive Bayes

- What if $\mathbf{x}$ is high-dimensional?
  - The $\mathbf{\Sigma}_k$ have $\mathcal{O}(D^2K)$ parameters, which can be a problem if D is large.
  - We already saw we can save some a factor of K by using a shared covariance for the classes.
  - Any other idea you can think of?
- **Naive Bayes**: Assumes features independent given the class

$$p(\mathbf{x} \,|\, t = k) = \prod_{j=1}^{D} p(x_j \,|\, t = k)$$

- Assuming likelihoods are Gaussian, how many parameters required for Naive Bayes classifier?
  - This is equivalent to assuming the x_j are uncorrelated, i.e. $\mathbf{\Sigma}$ is diagonal.
  - Hence, only D parameters for $\mathbf{\Sigma}$!

Gaussian Naïve Bayes

- Gaussian Naïve Bayes classifier assumes that the likelihoods are Gaussian:

$$p(x_j \,|\, t = k) = \frac{1}{\sqrt{2\pi}\sigma_{jk}} \exp\left[\frac{-(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right]$$

  (this is just a 1-dim Gaussian, one for each input dimension)

- Model the same as GDA with diagonal covariance matrix
- Maximum likelihood estimate of parameters

$$\begin{aligned}
\mu_{jk} &= \frac{\sum_{i=1}^{N} r_k^{(i)} \, x_j^{(i)}}{\sum_{i=1}^{N} r_k^{(i)}} \\
\sigma_{jk}^2 &= \frac{\sum_{i=1}^{N} r_k^{(i)} \, (x_j^{(i)} - \mu_{jk})^2}{\sum_{i=1}^{N} r_k^{(i)}} \\
r_k^{(i)} &= \mathbb{1}[t^{(i)} = k]
\end{aligned}$$

Decision Boundary: Isotropic

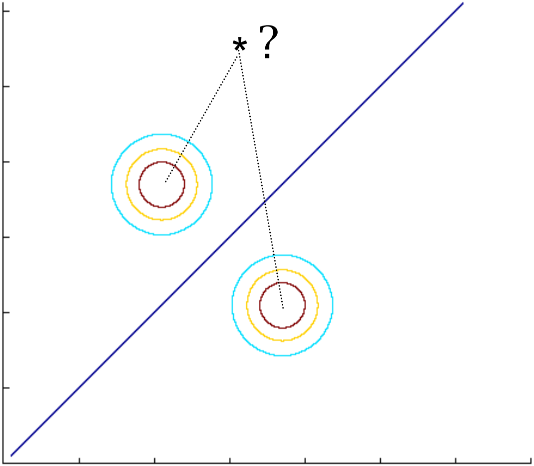
- We can go even further and assume the covariances are spherical, or isotropic.
- In this case: $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$ (just need one parameter!)
- Going back to the class posterior for GDA:

$$
\begin{aligned}
\log p(t_k|\mathbf{x}) &= \log p(\mathbf{x} \,|\, t_k) + \log p(t_k) - \log p(\mathbf{x}) \\
&= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_k^{-1}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \\
&\quad + \log p(t_k) - \log p(\mathbf{x})
\end{aligned}
$$

- Suppose for simplicity that $p(t)$ is uniform. Plugging in $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$ and simplifying a bit,

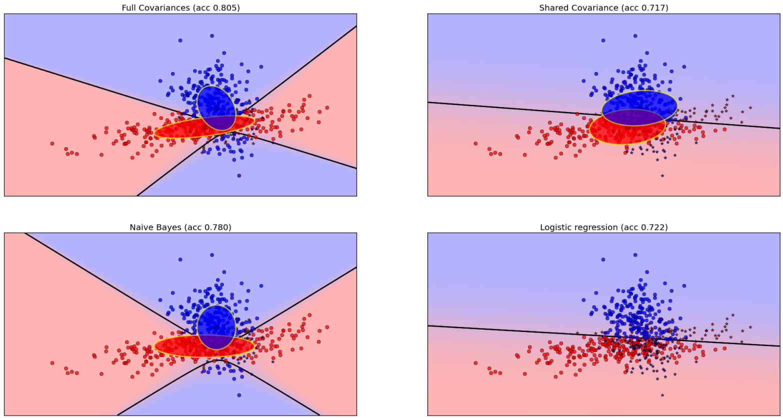
$$
\begin{aligned}
\log p(t_k \,|\, \mathbf{x}) - \log p(t_\ell \,|\, \mathbf{x}) &= -\frac{1}{2\sigma^2}\left[(\mathbf{x} - \boldsymbol{\mu}_k)^\top(\mathbf{x} - \boldsymbol{\mu}_k) - (\mathbf{x} - \boldsymbol{\mu}_\ell)^\top(\mathbf{x} - \boldsymbol{\mu}_\ell)\right] \\
&= -\frac{1}{2\sigma^2}\left[\|\mathbf{x} - \boldsymbol{\mu}_k\|^2 - \|\mathbf{x} - \boldsymbol{\mu}_\ell\|^2\right]
\end{aligned}
$$

Decision Boundary: Isotropic



- The decision boundary bisects the class means!

Example



Full Covariances (acc 0.805)
Shared Covariance (acc 0.717)
Naive Bayes (acc 0.780)
Logistic regression (acc 0.722)

Generative models - Recap

- GDA has quadratic (conic) decision boundary.
- With shared covariance, GDA is similar to logistic regression.
- Generative models:
  - Flexible models, easy to add/remove class.
  - Handle missing data naturally.
  - More "natural" way to think about things, but usually doesn't work as well.
- Tries to solve a hard problem (model $p(\mathbf{x})$) in order to solve a easy problem (model $p(t \mid \mathbf{x})$).