Introduction to Machine Learning

Chapter 7: Approaches to Classification

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

Department of Statistics – LMU Munich

CLASSIFICATION APPROACHES REMINDER

- **Discriminative models**: model $p(y|x)$ directly
  - Logistic/Softmax regression
  - kNN
- **Generative models**: model $p(x|y)$ and $p(y)$
  - Linear discriminant analysis (LDA)
  - Quadratic discriminant analysis (QDA)
  - Naïve Bayes

LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA follows a generative approach, each class density is modeled as a *multivariate Gaussian* with equal covariance, i. e. $\Sigma_k = \Sigma \quad \forall k$.

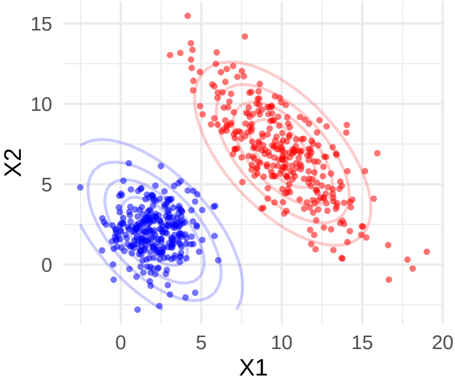
$$p(x|y = k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_k)^T\Sigma^{-1}(x - \mu_k)\right)$$

Parameters θ are estimated in a straight-forward manner by estimating

$$\begin{aligned}
\hat{\pi}_k &= n_k/n, \text{ where } n_k \text{ is the number of class } k \text{ observations} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i: y^{(i)} = k} x^{(i)} \\
\hat{\Sigma} &= \frac{1}{n - g} \sum_{k=1}^{g} \sum_{i: y^{(i)} = k} (x^{(i)} - \hat{\mu}_k)(x^{(i)} - \hat{\mu}_k)^T
\end{aligned}$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

- Each class fit as a Gaussian distribution over the feature space
- Different means but same covariance for all classes
- Rather restrictive model assumption.



LINEAR DISCRIMINANT ANALYSIS (LDA)

For the posterior probability of class k it follows:

$$
\begin{aligned}
\pi_k(x) &\propto \pi_k \cdot p(x|y=k) \\
&\propto \pi_k \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k\right) \\
&= \exp\left(\log \pi_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k\right)\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) \\
&= \exp\left(\theta_{0k} + x^T\theta_k\right)\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)
\end{aligned}
$$

by defining $\theta_{0k} := \log \pi_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$ and $\theta_k := \Sigma^{-1}\mu_k$.

LINEAR DISCRIMINANT ANALYSIS (LDA)

Finally, the posterior probability becomes

$$\pi_k(x) = \frac{\pi_k \cdot p(x|y=k)}{p(x)} = \frac{\exp(\theta_{0k} + x^T\theta_k)}{\sum_j \exp(\theta_{0j} + x^T\theta_j)}$$

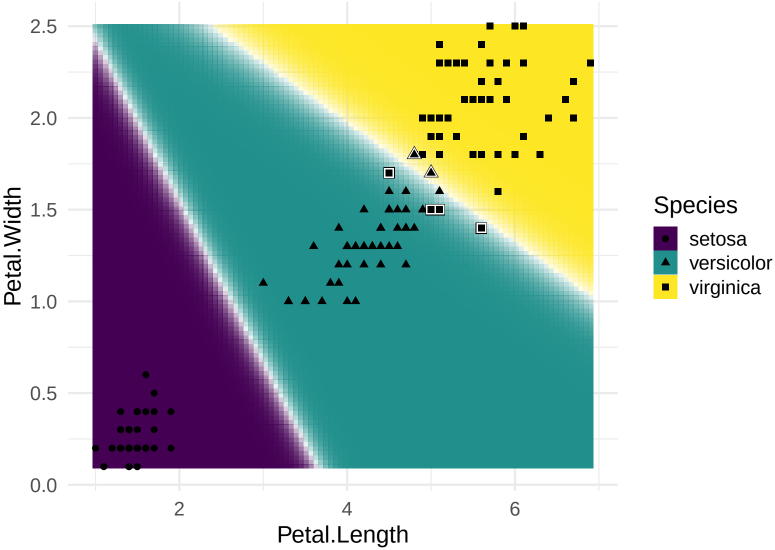
(the term $\exp(-\frac{1}{2}x^T\Sigma^{-1}x)$ will cancel out in numerator and denominator).

And (simplified) discriminant functions can be defined as

$$f_k(x) = \theta_{0k} + x^T\theta_k$$

Hence, LDA defines a *linear classifier* with linear decision boundaries.

LINEAR DISCRIMINANT ANALYSIS (LDA)



QUADRATIC DISCRIMINANT ANALYSIS (QDA)

QDA is a direct generalization of LDA, where the class densities are now Gaussians with unequal covariances Σ_k.

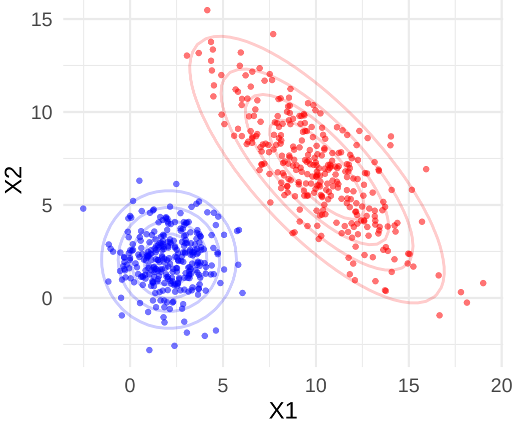
$$p(x|y = k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

Parameters are estimated in a straight-forward manner by:

$$\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n}, \text{ where } n_k \text{ is the number of class } k \text{ observations} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i: y^{(i)} = k} x^{(i)} \\
\hat{\Sigma}_k &= \frac{1}{n_k - 1} \sum_{i: y^{(i)} = k} (x^{(i)} - \hat{\mu}_k)(x^{(i)} - \hat{\mu}_k)^T
\end{aligned}$$

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

- Covariance matrices can differ over classes.
- Yields better data fit but also requires estimation of more parameters.



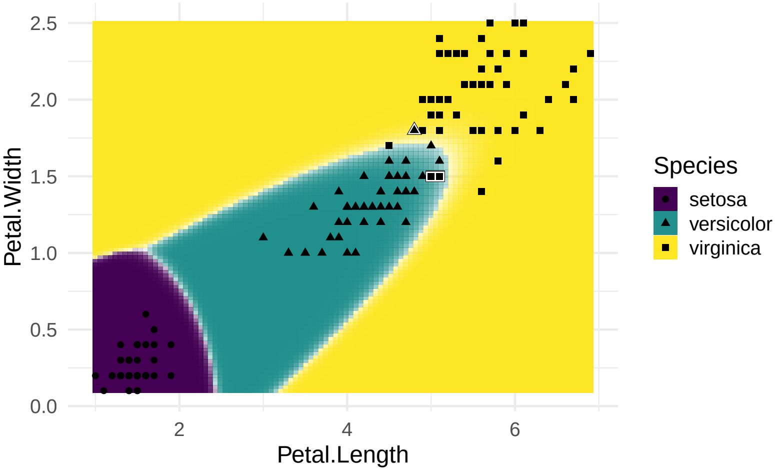
QUADRATIC DISCRIMINANT ANALYSIS (QDA)

$$
\begin{aligned}
\pi_k(x) &\propto \pi_k \cdot p(x|y=k) \\
&= \pi_k |\Sigma_k|^{-\frac{1}{2}} \exp(-\frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k)
\end{aligned}
$$

Taking the log of the above, we can define a discriminant function that is quadratic in x.

$$
\log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k
$$

QUADRATIC DISCRIMINANT ANALYSIS (QDA)



NAIVE BAYES CLASSIFIER

Another generative technique for categorical response $y \in \{1, \dots, g\}$ is called *Naive Bayes classifier*. Here, we make a "naive" *conditional independence assumption*: the features given the category y are conditionally independent of each other, so that we can simply write:

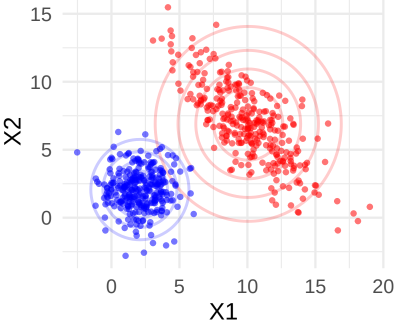
$$p(x|y=k) = p((x_1, x_2, ..., x_p)|y=k) = \prod_{j=1}^{p} p(x_j|y=k).$$

Putting this together we get

$$\pi_k(x) \propto \pi_k \cdot \prod_{j=1}^{p} p(x_j|y=k)$$

NAIVE BAYES CLASSIFIER

- Covariance matrices can differ over both classes but assumed to be diagonal.
- Assumption of uncorrelated features (!!)
- Often performs well despite this usually wrong assumption
- Easy to deal with mixed features (metric and categorical)



NAIVE BAYES CLASSIFIER

Parameters estimation now has become simple, as we only have to estimate $p(x_j|y = k)$, which is univariate (given the class k).

For numerical x_j, often a univariate Gaussian is assumed, and we estimate (μ_j, σ_j^2) in the standard manner. Note, that we now have constructed a QDA model with strictly diagonal covariance structures for each class, hence this leads to quadratic discriminant functions.

For categorical features x_j, we simply use a Bernoulli / categorical distribution model for $p(x_j|y = k)$ and estimate the probabilities for (j, k) by simply counting of relative frequencies in the standard manner. The resulting classifier is linear in these frequencies.

NAIVE BAYES CLASSIFIER