itsl-401

Generative models for classification

Suppose Y can take on K distinct class values. Let π_k represent the overall (prior) probability that a randomly chosen observation comes from the k^{th} class. Let $f_k(X) := \Pr(X|Y = k)$ denote the density function of X for an observation that comes from the k^{th} class. Then, the Bayes' theorem states {

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

}

Back:

Classification –

itsl-402

Linear discrimination analysis for $p = 1$

It is assumed that $f_k(x) := \mathsf{Pr}(X = x|Y = k)$ is normal. In the one-dimensional setting, the normal density takes the form {

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

} where μ_k and σ_k^2 are the {mean and variance parameters for the k^{th} class.}

Back:

Classification–

itsl-403

Linear discrimination analysis for $p = 1$

The Bayes Classifier involves {assigning an observation $X = x$ to the class for which $p_k(x)$ is largest}. Assuming $f_k(x)$ is normal and that there is a shared variance term across all K classes, this is equivalent to assigning the observation to the class for which {

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

} is largest.

Back:

itsl-404

Linear discrimination analysis for $p = 1$

The linear discriminant analysis (LDA) method approximates the Bayes classifier by using estimates for $\hat{\pi}_k$, $\hat{\mu}_k$ and $\hat{\sigma}^2$, given by,

{

$$\begin{aligned}
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \\
\hat{\pi}_k &= \frac{n_k}{n},
\end{aligned}$$

}

where n is the total number of training observations and n_k is the number in the k^{th} class.

Back:

Linear discrimination analysis for $p = 1$

The LDA classifier results from assuming that the observations {within each class come from a normal distribution} with a {class specific mean and common variance σ^2}, and plugging esimates for those parameters into the {Bayes classifier}.

Back:

itsl-406

Classification and diagnostic testing

False positive rate	{FP / N }	{Type I error, 1 - Specificity }
True positive rate	{TP / P }	{1 - Type II Error, power, sensitivity, recall }
Positive predictive value	{TP / $\hat{P}$ }	{Precision, 1 - false discovery proportion }
Negative predictive value	{TN / $\hat{N}$ }	

Back:

Classification

itsl-407

Quadratic discriminant analysis

What is the difference between QDA and LDA?
QDA assumes that an observation from the k^{th} class is of the form $\{X \sim \mathcal{N}(\mu_k, \Sigma_k)\}$, where $\{\Sigma_k$ is a covariance matrix for the k^{th} class$\}$. LDA assumes that all observations $\{$share a common covariance matrix $\Sigma\}$.

Back:

Naive Bayes

The naive Bayes classifier makes the assumption that, {

Within the k^{th} class, the p predictors are independent.

} Mathematically, this assumption means, {

$$f_k(x) = f_{k1}(x_1) \times f_{k2}x_2 \times \cdots \times f_{kp}(x_p),$$

} where {f_{kj} is the density function of the j^{th} predictor among observations in the k^{th} class}.

Back:

itsl-409

Naive Bayes

With the naive Bayes assumption (that the p covariates are independent within each class), and using Bayes theorem, we get posterior probability, {

$$\mathsf{Pr}(Y = k|X = x) = \frac{\pi_k \times f_{k1}(x_1) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times \cdots \times f_{lp}(x_p)},$$

} for $k = 1, \ldots, K$.

Back:

itsl-410

Naive Bayes

Three options to estimate the one-dimensional density function f_{kj} using training data $x_{1j}, \ldots, x_{pj}$ are,

- X_j quantitative: {
  - Can assume $X_j|Y \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$ (this is QDA with an additional assumption that the class-specific covariance matrix is diagonal);
  - Non-parametric estimate using the histogram or kernel density function.

  }
- X_j qualitative: {count the proportion of training observations for the j^{th} predictor corresponding to each class}.

Back:

Classification

itsl-411

KNN vs LDA and QDA

- KNN is {non-parametric} and therefore we expect it to dominate LDA and logistic regression when {the decision boundary is highly non-linear, provided n is large and p is small}.

- Where the decision boundary is non-lienar but n is modest or p is not very small, then {QDA may be preferred to KNN}. This is because {QDA can prove a non-linear decision boundary while taking advantage of a parametric form}.

- Unlike logistic regression, KNN does not {tell us which predictors are important}.

Back: