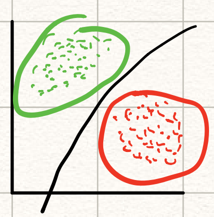
Lecture 7, Mar 4 2021

More on Gaussian Class Conditionals (Generative Model)

Ex. what model looks like?



The dist. used is mutivariate normal. We need:

- mean vector $\in \mathbb{R}^D$
- covariance vector $\in \mathbb{R}^{D \times D}$
- priors (% of data in each class)

We can also regularize.

Naive Bayes //

This model assume the features in the data are independent. This allows us to reduce the dimensionality of high dimensional data with a less expressive model.

Instead of estimating one d-dimensional density, we estimate d one-dimensional densities.

$$p(x|c) = \prod_{i=1}^{d} p(x_i|c)$$

We will use discrete input vectors instead of real vectors like before.

Ex. Document Vectors

Vector is binary with preset labels for each index.

$$a[i] = \begin{cases} 1 & \text{if document contains label} \\ 0 & \text{if document doesn't contain label} \end{cases}$$

Since most indices are going to be 0, we use Naive Baye's to reduce the dimensionality

GCC would be impossible.

You can regularize the learning process when you don't have enough data. When data is low

$$p(x|c) = p(x_1|c) \times \cdots \times \underline{\underline{0}} \times \cdots p(x_d|c) = 0$$

Thus we regularize by:

$$b_j = \frac{N_j}{N} \Rightarrow \frac{N_j + \beta}{N + k\beta}$$

$$a_{ij} = \frac{N_{ij}}{N_j} \Rightarrow \frac{N_{ij} + \alpha}{N_j + 2\alpha}$$

Generalization//

Is the model going to preform well on new data?

To make sure the model generalizes, we partition
- training data : used to create the initial
- validation data : used to pick optimal hyper-param
- testing data : used to test generalization

We choose best hyper-param using <u>grid search</u>.
Create matrix of different combinations of hyperparams, choose the best performing one on the validation data.
- if dim of grid to big, do random search, choose from random subset.

Bias vs Variance

Bias

The square of the best predictor of x, $f(x)$ and the average predictor of x, $h(x)$ from multiple models.

Variance

The square of the average predictor of x, $h(x)$ from multiple models and a predictor of x from a particular model.

$$E_{D,\varepsilon}((f-\hat{y})^2)$$

$$= \underbrace{E((f-h)^2)}_{\text{bias}} + \underbrace{E((h-\hat{y})^2)}_{\text{variance}}$$

Underfit data: high bias, low variance
(low params)

Overfit data: low bias, high variance
(many params)

Overfit Data II
Accidental Regularities
- when the model perceives something insignificant as significant.
    - model thinks all shoes must have nike logo since, data comes from nike.