

# 1 Let's Generative Some Definitions!

**Generative Models:** class of models that can generate new data points that are similar to the training data. That is, generative models model  $P(Y | \theta_{\text{class}})$  and  $P(X | Y, \theta_{\text{class conditional}})$ . In addition to generating new data, generative models can also be used for prediction.

The parameters  $\theta_{\text{class}}$  and  $\theta_{\text{class conditional}}$  are learned from the data. Bayes rule can then be used to compute  $P(Y | X, \theta)$

$$P(Y | X, \theta_{\text{class conditional}}, \theta_{\text{class}}) \propto P(X | Y, \theta_{\text{class conditional}}) P(Y | \theta_{\text{class}})$$

Properties:

- More model assumptions involved
- Rich model allows for generating new data points that are similar to the training data.
- Can be used for unsupervised learning, where the class labels are not known.

**Discriminative Models:** Discriminative models model  $P(Y | X, \theta)$  directly. The parameters,  $\theta$ , are also learned from the data.

Properties:

- Less model assumptions needed
- Requires more labeled data

## 1.1 Generative + MAP

We can also combine generative models with MAP by adding a prior on the parameters:

$$P(Y | X, \theta) P(\theta) \propto P(X | Y, \theta_{\text{class conditional}}) P(Y | \theta_{\text{class}}) P(\theta_{\text{class conditional}}) P(\theta_{\text{class}})$$

## 1.2 Naive Bayes Recap

Naive Bayes is a probabilistic model that simplifies a generative model by using the naive Bayes assumption, which states that all input features of a data point are conditionally independent from each other given the output value. That is, for all  $i \neq j$ ,  $X_i$  and  $X_j$  are conditionally independent given  $Y$  or:

$$P(X_1, X_2, \dots, X_j | Y) = \prod_{j=1}^M P(X_j | Y)$$

**Naive Bayes Optimization:**

1. Similar to MLE, estimate the parameters for the class prior distribution,  $P(Y | \theta_{\text{class}})$ .
2. **For each class label  $y$** , use the data points with label  $y$  only to estimate the parameters (similar to MLE) for  $P(X_j | Y = y, \theta_{\text{class conditional}, y, j})$  **for each feature  $X_j$ , independently**.

**Naive Bayes Inference (Prediction):** Given a new set of input features  $x_1, \dots, x_M$ , class label with the highest posterior probability,  $P(Y = y | x_1, \dots, x_M)$  by taking advantage of Bayes theorem:

$$\underset{y}{\operatorname{argmax}} P(Y = y) \prod_{j=1}^M P(X_j | Y = y)$$

## 2 Comparing models

|                       | MLE | MAP |
|-----------------------|-----|-----|
| <b>Discriminative</b> |     |     |
| <b>Generative</b>     |     |     |

Consider the list of models we've learned so far and place them in the correct boxes above:

- Linear regression
- Logistic regression
- Linear regression with L2 regularization
- Logistic regression with Laplace prior
- Logistic regression with polynomial features
- Naive Bayes

### 2.1 Reminder: Regularization and MAP

|                         | Regularization<br>Penalty | Prior                             |
|-------------------------|---------------------------|-----------------------------------|
| <b>Ridge Regression</b> | $\ \mathbf{w}\ _2^2$      | $w_j \sim \mathcal{N}(0, \tau^2)$ |
| <b>Lasso</b>            | $\ \mathbf{w}\ _1$        | $w_j \sim \text{Laplace}(0, b)$   |

### 3 Gaussian Discriminant Analysis (A generative method)

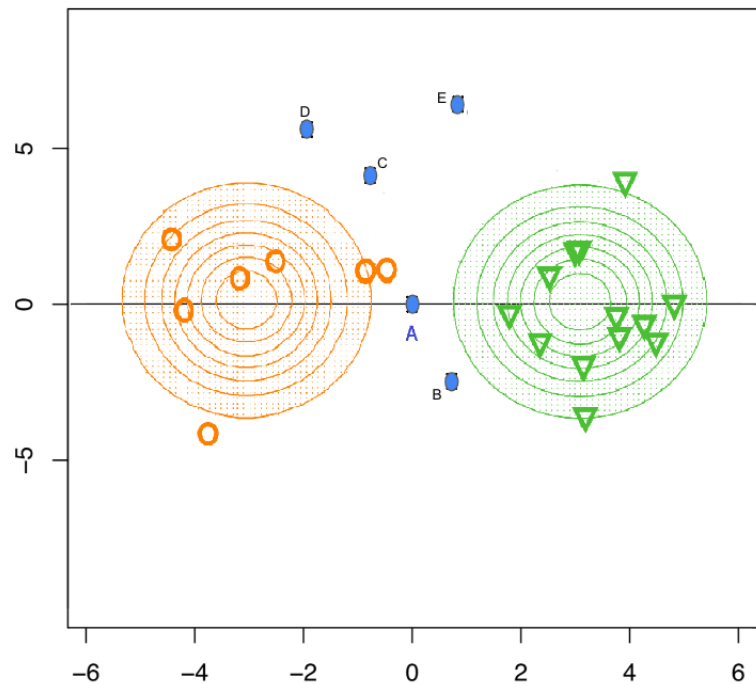
Gaussian discriminant analysis is used when the input features are continuous and  $p(\mathbf{x} | y)$  is modeled as a multivariate Gaussian distribution.

Note: Since we're dealing with  $p(\mathbf{x} | y)p(y)$ , it is a generative model, despite its name!

- a. Consider two Gaussian distributions as formulated and visualized below:

$$\mathbf{x} | y = 0 \sim \mathcal{N}(\mu_{y=0}, \mathbf{I})$$

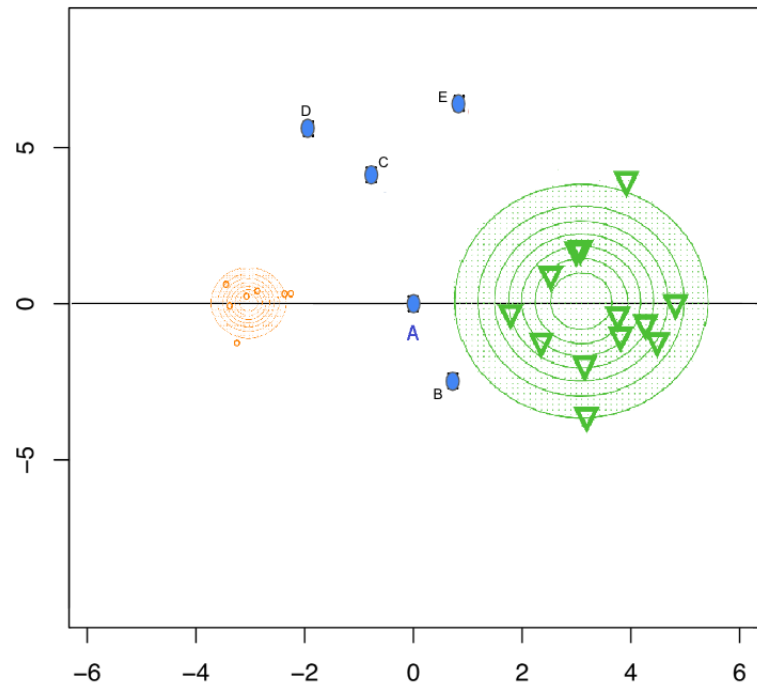
$$\mathbf{x} | y = 1 \sim \mathcal{N}(\mu_{y=1}, \mathbf{I})$$



- i.  $X$  are some observed data points. You are given that point  $A$  lies on the midpoint between the two distribution centers. Label each data point with its likely class.  
(hint: Both distributions have the same covariance!)

- ii. If we were to draw a boundary separating the two distributions, where would the boundary be and what would it look like?

- b. Now supposed a different distribution, whose covariance matrix is  $\frac{1}{5}I$ . Re-label the data points again. Point  $A$  lies on the midpoint between the two distribution centers. How does the boundary change?



## 4 Naive Bayes with Laplace Smoothing

We're going to repeat the example from last week but introduce Laplace Smoothing as the Prior. It was very convenient to assume that the features are independent given the class—it allowed us to simplify the calculations a lot. But, we found a problem with our example last week. Whenever a word/token didn't occur for a given class, the probability for that token given that class becomes 0, hence the probability of that class also becomes 0. This is a massive drawback. We cannot extrapolate our model to an example that contains a word that we have not seen before in our training data. If we tried, the probability for each class would just be 0.

Laplace smoothing, is a technique used to address this issue by adding a small pseudo-count (denoted by  $\alpha$ ) to each observation. This ensures that no probability estimation becomes zero.

To be more precise,

$$P(x_i = k|Y = y) = \frac{(\text{Number of samples that belong to class } y \text{ and in which } x_i \text{ takes on class } k) + \alpha}{\text{Number of samples that belong to class } y + \alpha K}$$

$K$  is the total number of classes that  $x_i$  can take on and  $\alpha$  is a constant hyperparameter. Compare this to the formula from before:

$$P(x_i = k|Y = y) = \frac{\text{Number of samples that belong to class } y \text{ and in which } x_i \text{ takes on class } k}{\text{Number of samples that belong to class } y}$$

**Note:** It's important to clarify that "Laplace" smoothing is not related to the Laplace distribution. The term "Laplace" here simply refers to the technique of smoothing probabilities with pseudocounts.

For the case of Bernoulli distributions, Laplace smoothing is equivalent to using a  $\text{Beta}(\alpha + 1, \alpha + 1)$  prior. For categorical distributions, Laplace smoothing can be interpreted as using a Dirichlet prior.

Let's repeat the same example from last week but use Laplace Smoothing this time to see how we can avoid getting zero as our class probability. Use  $\alpha = 1$ .

Consider the following training samples:

| SPAM | Email Body         |
|------|--------------------|
| 1    | Money is free now  |
| 0    | Pat teach 315      |
| 0    | Pat free to teach  |
| 1    | Sir money to teach |
| 1    | Pat free money now |
| 0    | Teach 315 now      |
| 0    | Pat to teach 315   |

The vocabulary consists of the following words: {315, free, is, money, now, Pat, Sir, teach, to, tomorrow}. Compute the following probabilities:

1. Fill in the tables below:

|            |            |
|------------|------------|
| $P(Y = 1)$ | $P(Y = 0)$ |
|            |            |

| $j$      | $P(X_j = 1 \mid Y = 1)$ | $P(X_j = 1 \mid Y = 0)$ |
|----------|-------------------------|-------------------------|
| 315      |                         |                         |
| free     |                         |                         |
| is       |                         |                         |
| money    |                         |                         |
| now      |                         |                         |
| Pat      |                         |                         |
| Sir      |                         |                         |
| teach    |                         |                         |
| to       |                         |                         |
| tomorrow |                         |                         |

2. Consider the following email body:  $X = \text{Pat teach now}$ . Accounting for the laplace smoothing, fill in the table below:

| $j$      | $x$ | $P(X_j = x \mid Y = 1)$ | $P(X_j = x \mid Y = 0)$ |
|----------|-----|-------------------------|-------------------------|
| 315      | 0   |                         |                         |
| free     | 0   |                         |                         |
| is       | 0   |                         |                         |
| money    | 0   |                         |                         |
| now      | 1   |                         |                         |
| Pat      | 1   |                         |                         |
| Sir      | 0   |                         |                         |
| teach    | 1   |                         |                         |
| to       | 0   |                         |                         |
| tomorrow | 0   |                         |                         |

3. Reminder that with naive Bayes,  $P(Y, X_1, \dots, X_M) = \prod P(X \mid Y)P(Y)$ .

|                             |                             |
|-----------------------------|-----------------------------|
| $P(Y = 1, X_1, \dots, X_M)$ | $P(Y = 0, X_1, \dots, X_M)$ |
|                             |                             |

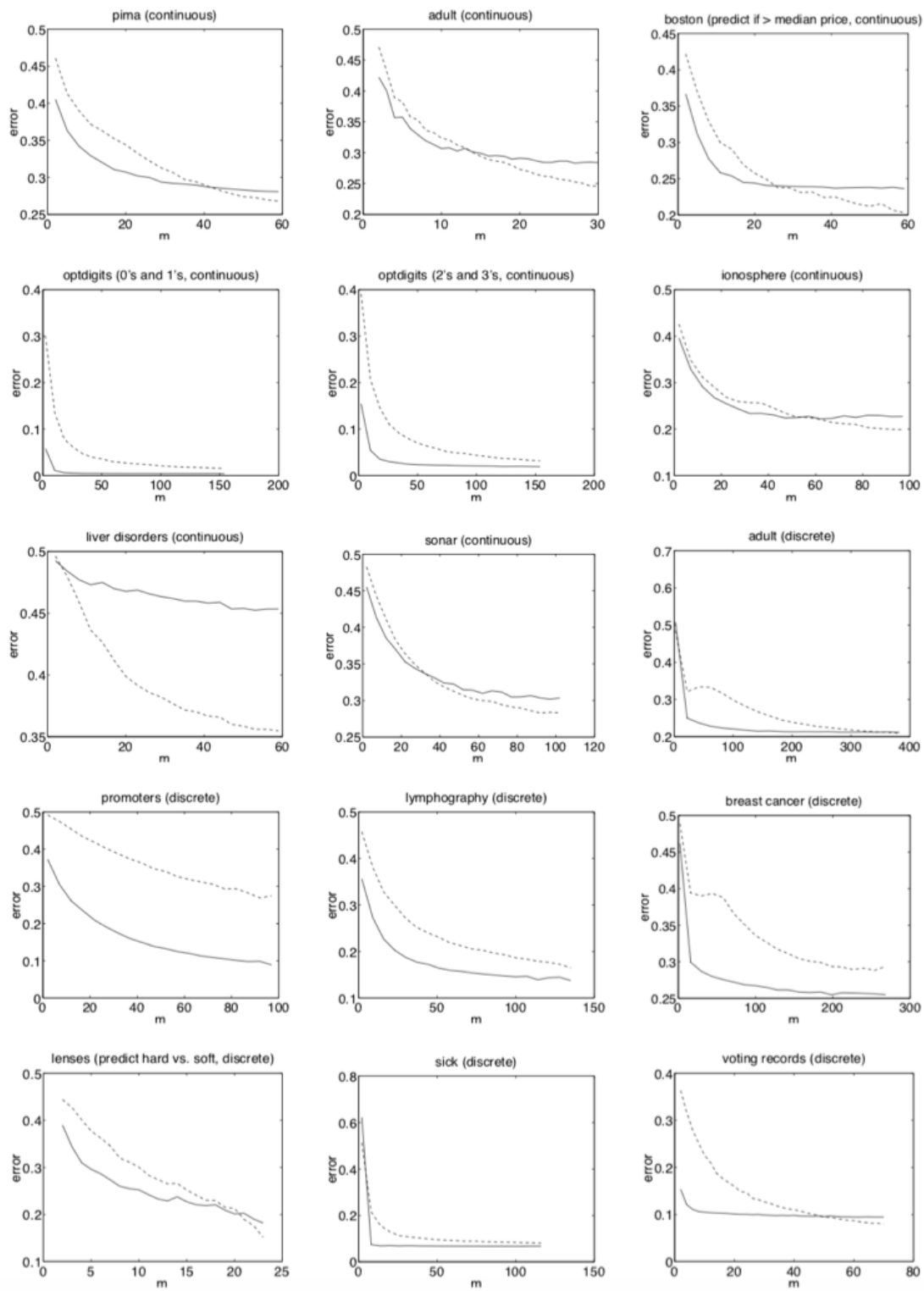
4.

|                                 |                                 |
|---------------------------------|---------------------------------|
| $P(Y = 1 \mid X_1, \dots, X_M)$ | $P(Y = 0 \mid X_1, \dots, X_M)$ |
|                                 |                                 |

We see that  $P(Y = 1 \mid X_1, \dots, X_M)$  is no longer 0.

What's the effect of  $\alpha$ ? Compare when  $\alpha = 1$  with when  $\alpha = \infty$

## 5 Performance of Generative vs Discriminative models



These are 15 experiments run by Andrew Ng and Michael Jordan at U.C. Berkeley using 15 standard UCI data sets, in order to compare the performance of Generative vs Discriminative models in real applications. (FYI, this is the link to the original paper: <https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf>) In each of the plots, the X-axis is number of samples, and Y-axis is average error across 1,000 random splits of training/validation set. Look at the graphs and answer the following:

- a. The two models considered are Naive Bayes for Generative model and Logistic Regression for Discriminative model, where the dashed line represents Logistic Regression and the solid line represents Naive Bayes. According to these results, which model has the better asymptotic performance?

- b. What performance advantages does each type of model possess?