THE UNIVERSITY OF CHICAGO

STAT 37710 / CMSC 35400 / CAAM 37710
Machine Learning

Generative Models for Classification

Cong Ma

Discriminative modeling

- Discriminative models aim to estimate conditional distribution

$$P(y \mid \mathbf{x})$$

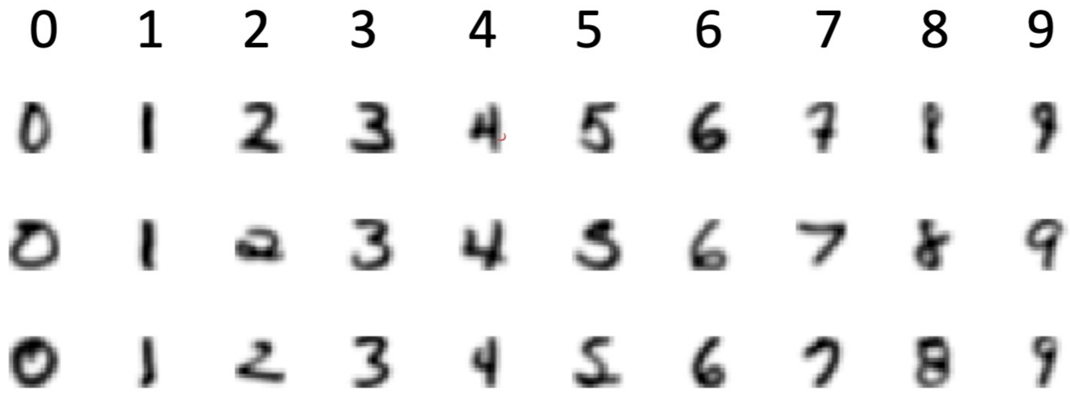
- Generative models aim to estimate joint distribution

$$P(y, \mathbf{x})$$

- Can derive conditional from joint distribution, but not vice versa.

Typical approaches to generative modeling

- Estimate prior on labels P(y)
- Estimate conditional distribution P(**x** | y) for each class y
- Obtain predictive distribution using Bayes' rule: P(y | **x**) = P(y) P(**x** | y) / Z



Example: hand-written digits

Naïve Bayes classifier (NB)

- Model class label as generated from categorical variable

$$P(Y = y) = p_y, \;\; y \in \mathcal{Y} = \{1, \ldots, c\}$$

- Model features as conditionally independent given label

$$P(X_{[1]}, \ldots, X_{[d]} \mid Y) = \prod_{i=1}^{d} P(X_{[i]} \mid Y)$$

  - given class label, each feature is generated independently of the other features
  - need to specify feature distribution $P(X_{[i]} \mid Y)$

Gaussian Naïve Bayes classifier (GNB)

- Model class label as generated from categorical variable

$$P(Y = y) = p_y, \ \ y \in \mathcal{Y} = \{1, \ldots, c\}$$

- Model features as conditionally independent Gaussians

$$P(X_{[1]}, \ldots, X_{[d]} \mid Y) = \prod_{i=1}^{d} P(X_{[i]} \mid Y)$$

$$P(x_{[i]} \mid y) = \mathcal{N}(x_{[i]} \mid \mu_{y,[i]}, \sigma^2_{y,[i]})$$

- How do we estimate the parameters?

MLE for P(y)

$$\mathcal{Y} = \{-1, +1\} \qquad P(Y = +1) = p \qquad D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

- Estimate P(y) using *D* via MLE:

$$\max_p P(D \mid p) = \prod_{i=1}^{n} p^{[y_i=1]}(1-p)^{[y_i=-1]} = p^{n_+}(1-p)^{n_-}$$

where n+ (resp. n−) corresponds to the number of + (resp. -) instances in D.

- The log-likelihood is $\log P(D \mid p) = n_+ \log p + n_- \log(1-p)$
- Taking the gradient and set to 0, we get MLE for label distribution:

MLE for P(x|y)

$$P(x_{[i]} \mid y) = \mathcal{N}(x_{[i]}; \mu_{y,[i]}, \sigma^2_{y,[i]}) \qquad D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

- MLE for feature distribution:

Decision rules

- We have estimated P(y) and P(x | y). In order to predict label y for a new data point **x**, use the Bayes' rule

$$P(y \mid \mathbf{x}) = \frac{1}{Z} P(y) P(\mathbf{x} \mid y), \quad \text{where } Z = \sum_y P(y) P(\mathbf{x} \mid y)$$

- To minimize misclassification error, predict:

Gaussian Naive Bayes classifiers

- MLE for class label distribution $\hat{P}(Y=y)=\hat{p}_y=\dfrac{\mathrm{Count}(Y=y)}{n}$

- MLE for feature distribution: $\hat{P}(x_{[i]} \mid y)=\mathcal{N}(x_{[i]};\hat{\mu}_{y[i]},\sigma^2_{y[i]})$

$$\hat{\mu}_{y[i]}=\frac{1}{\mathrm{Count}(Y=y)}\sum_{j:y_j=y}x_{j[i]}$$

$$\hat{\sigma}^2_{y[i]}=\frac{1}{\mathrm{Count}(Y=y)}\sum_{j:y_j=y}\left(x_{j[i]}-\hat{\mu}_{y[i]}\right)^2$$

- Prediction given new point **x**:

$$y=\arg\max_{y_j}\hat{P}(y_j \mid \mathbf{x})=\arg\max_{y_j}\hat{P}(y_j)\prod_{i=1}^{d}\hat{P}(x_{[i]} \mid y_j)$$

Example: decision boundary (1D)

- Assume $d = 1, \mathbf{x} = x_{[1]}, \mathcal{Y} = \{-1, +1\}$ and $P(Y = +1) = 0.5$

- The decision boundary for a new point **x** is

$$y = \arg\max_{y_j} P(y_j \mid \mathbf{x}) = \arg\max_{y_j} \hat{P}(y_j)\hat{P}(\mathbf{x} \mid y_j)$$
$$= \arg\max_{y_j} P(\mathbf{x} \mid y_j)$$

Decision rules for binary classification

- We want to predict $y = \arg\max_{y_j} \hat{P}(y_j \mid \mathbf{x}) = \arg\max_{y_j} \hat{P}(y_j) \prod_{i=1}^{d} \hat{P}(x_{[i]} \mid y_j)$

- For binary tasks (i.e. c = 2, y ∈ {−1, +1}), this is equivalent to

$$y = \text{sign}\underbrace{\left(\log \frac{P(Y = +1 \mid \mathbf{x})}{P(Y = -1 \mid \mathbf{x})}\right)}_{f(\mathbf{x})}$$

- Discriminant function

  - The function $f(\mathbf{x}) = \log \frac{P(Y = +1 \mid \mathbf{x})}{P(Y = -1 \mid \mathbf{x})}$ is called discriminant function.

Example: GNB (c=2, class-invariant variance)

- Assume
  - Binary classes: $\mathcal{Y} = \{-1, +1\}$
  - Class independent variance: $P(\mathbf{x} \mid y) = \prod_i \mathcal{N}(x_{[i]}; \mu_{y,[i]}, \sigma^2_{[i]})$
- Then $f(\mathbf{x}) = \log \dfrac{P(Y = +1 \mid \mathbf{x})}{P(Y = -1 \mid \mathbf{x})} = \mathbf{w}^\top \mathbf{x} + b$

where $w_{[i]} = \dfrac{\hat{\mu}_{+,[i]} - \hat{\mu}_{-,[i]}}{\hat{\sigma}^2_{[i]}}, \quad b = \log \dfrac{\hat{p}_+}{1 - \hat{p}_+} + \sum_{i=1}^{d} \dfrac{\mu^2_{-,[i]} - \mu^2_{+,[i]}}{2\hat{\sigma}^2_{[i]}}$

How?

$$f(\mathbf{x}) = \log \frac{P(Y=+1 \mid \mathbf{x})}{P(Y=-1 \mid \mathbf{x})} = \log \frac{P(Y=+1)\prod_{i=1}^{d} P(x_{[i]} \mid Y=+1)/P(\mathbf{x})}{P(Y=-1)\prod_{i=1}^{d} P(x_{[i]} \mid Y=-1)/P(\mathbf{x})}$$

$$= \log \frac{\hat{p}_+}{1-\hat{p}_+} + \log \prod_{i=1}^{d} \frac{P(x_{[i]} \mid Y=+1)}{P(x_{[i]} \mid Y=-1)}$$

$$= \log \frac{\hat{p}_+}{1-\hat{p}_+} + \log \prod_{i=1}^{d} \frac{\frac{1}{\sqrt{2\pi}\sigma_{[i]}} \exp\left(-\frac{1}{2\sigma_{[i]}^2}\left(x_{[i]} - \mu_{+1,[i]}\right)^2\right)}{\frac{1}{\sqrt{2\pi}\sigma_{[i]}} \exp\left(-\frac{1}{2\sigma_{[i]}^2}\left(x_{[i]} - \mu_{-1,[i]}\right)^2\right)}$$

$$= \log \frac{\hat{p}_+}{1-\hat{p}_+} + \sum_{i=1}^{d} \left(-\frac{1}{2\sigma_{[i]}^2}\left(x_{[i]} - \mu_{+1,[i]}\right)^2 + \frac{1}{2\sigma_{[i]}^2}\left(x_{[i]} - \mu_{-1,[i]}\right)^2\right)$$

$$= \sum_{i=1}^{d} \underbrace{\left(\frac{\hat{\mu}_{+,[i]} - \hat{\mu}_{-,[i]}}{\hat{\sigma}_{[i]}^2}\right)}_{w_{[i]}} x_{[i]} + \underbrace{\log \frac{\hat{p}_+}{1-\hat{p}_+} + \sum_{i=1}^{d} \frac{\mu_{-,[i]}^2 - \mu_{+,[i]}^2}{2\hat{\sigma}_{[i]}^2}}_{b}$$

Gaussian NB (c=2): f vs. class probability

$$f(\mathbf{x}) = \log \frac{P(Y=+1 \mid \mathbf{x})}{P(Y=-1 \mid \mathbf{x})}$$

$$\Leftrightarrow \quad P(Y=+1 \mid \mathbf{x}) = \frac{1}{1+\exp\left(-f(\mathbf{x})\right)} = \sigma(f(\mathbf{x}))$$

- Therefore, for 2-class GNB with class independent variance is

$$P(Y=+1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

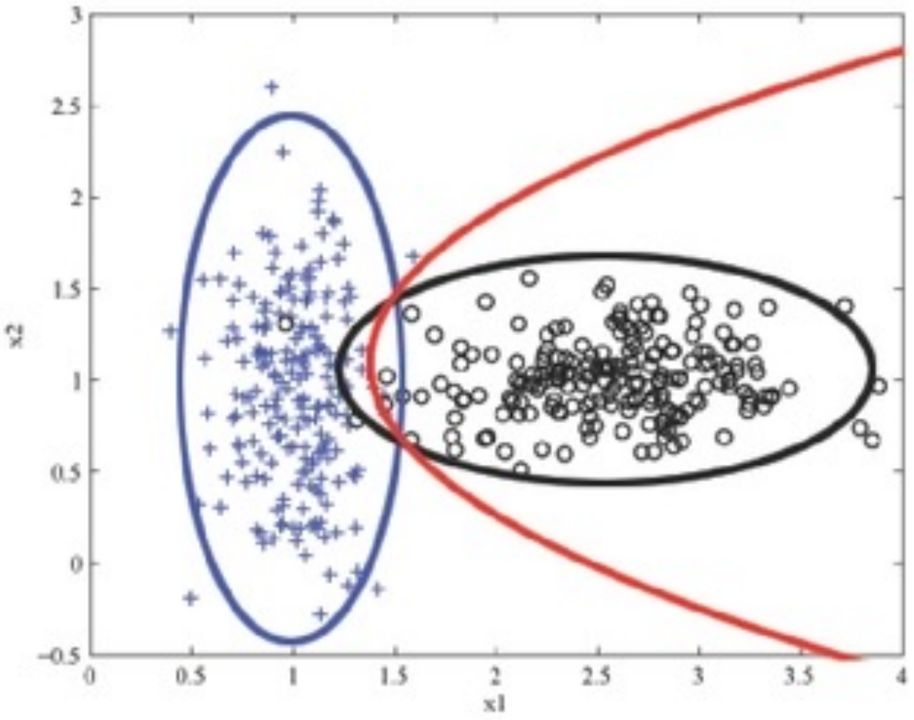
This is of the same form as logistic regression.

- If model assumptions are met, GNB will make same predictions as Logistic Regression!

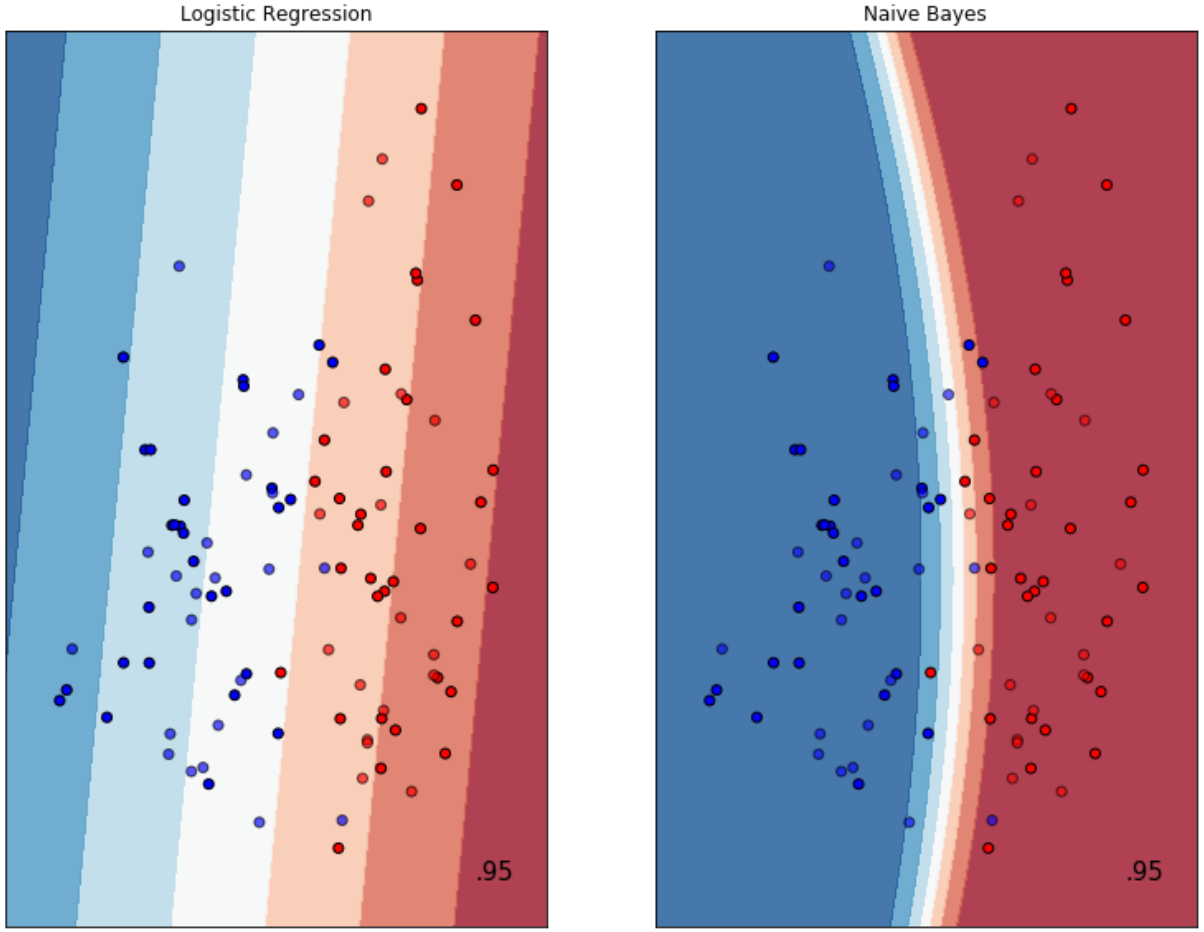
Gaussian NB (c=2): Decision boundary

- Our analysis on the previous slide is for
  - binary classification
  - class independent variance

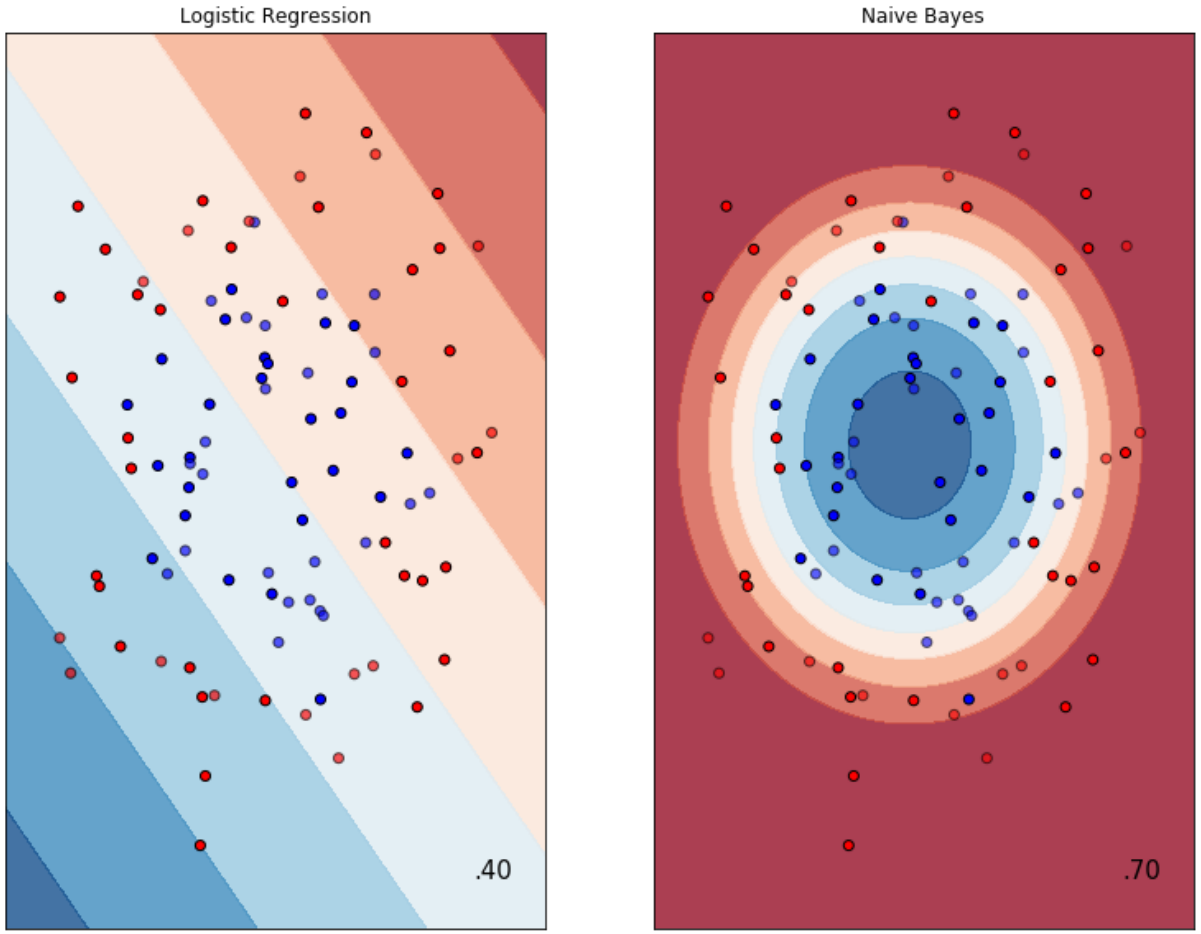
- Nevertheless, one can still apply GNB to datasets violating these assumptions
  - e.g., multi-class, arbitrary variance



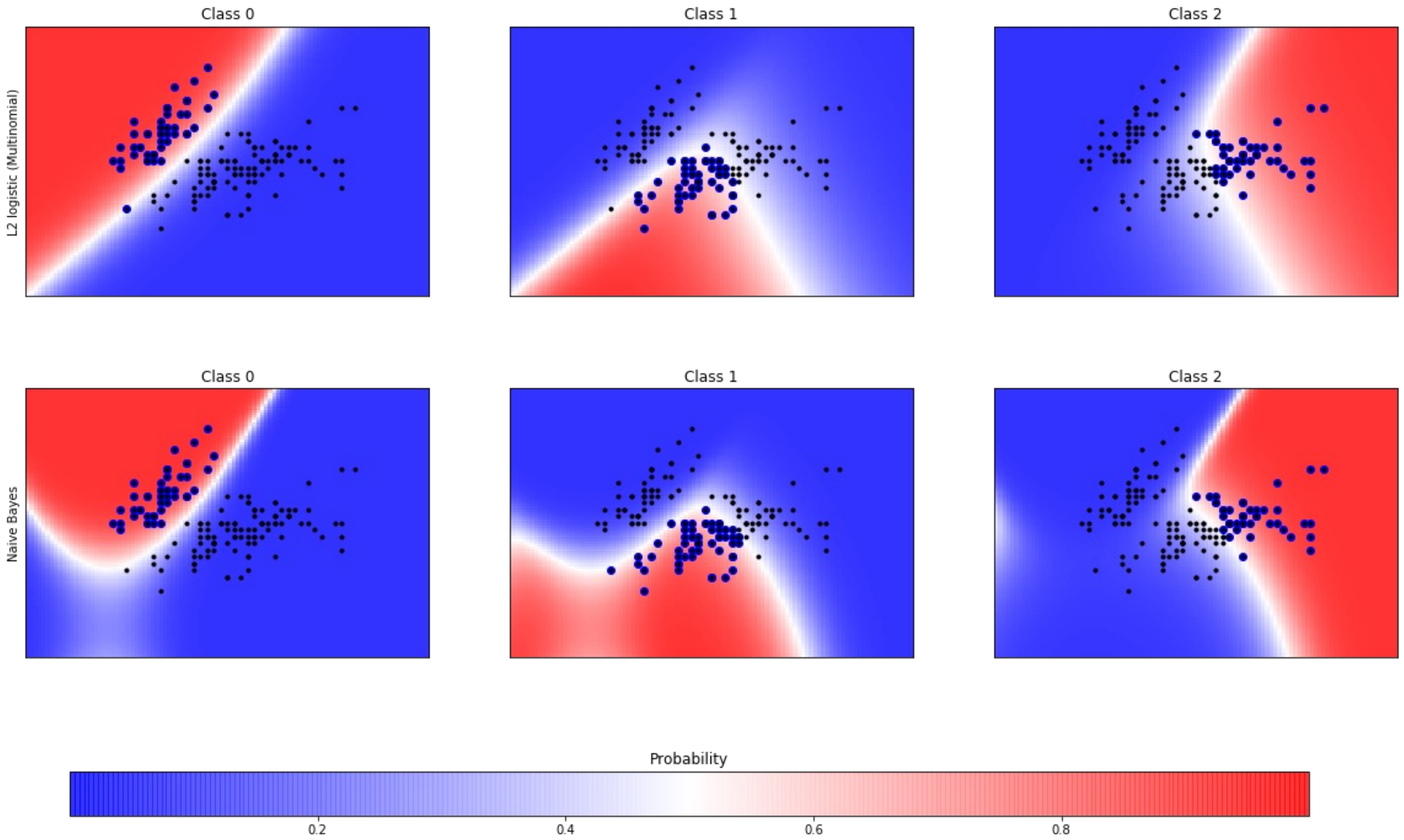
Demo: Gaussian NB vs LR (linear)



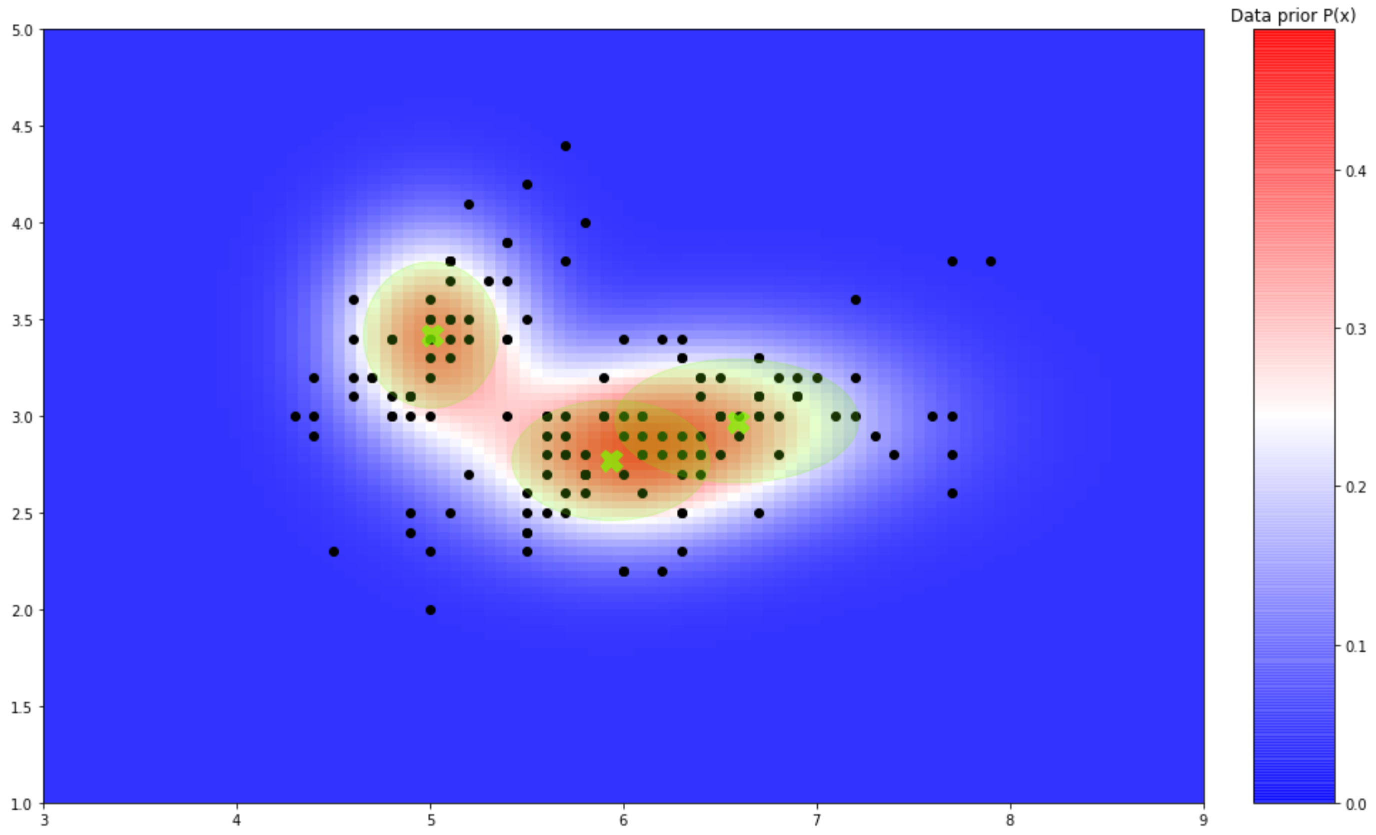
Gaussian NB vs. LR (circle)



Gaussian NB vs LR (multi-class)



Gaussian NB (data likelihood): Anomaly detection



Limitation of Naïve Bayes models

- Assume $\mathcal{Y} = \{-1, +1\}$ $\quad P(Y = +1) = 0.5 \qquad x_{[1]} = \cdots = x_{[d]}, \forall \mathbf{x} \in \mathcal{X}$

  $P(X_{[i]} = x \mid y) = \mathcal{N}(x \mid \mu_y, 1)$

- We consider the discriminant function for two GNB variants:
  - For GNB that only uses $X_{[1]} : f_1(\mathbf{X}) = \log \dfrac{P(Y = +1 \mid X_{[1]} = x)}{P(Y = -1 \mid X_{[1]} = x)}$
  - For GNB that uses $X_{[1]}, \ldots, X_{[d]} \; : \; f_2(\mathbf{X}) = \log \dfrac{P(Y = +1 \mid X_{[1]} = x, \ldots, X_{[d]} = x)}{P(Y = -1 \mid X_{[1]} = x, \ldots, X_{[d]} = x))}$

$$= \log \frac{P(X_{[1]} = x, \ldots, X_{[d]} = x \mid Y = +1)}{P(X_{[1]} = x, \ldots, X_{[d]} = x \mid Y = -1))}$$

$$= \log \prod_{i=1}^{d} \frac{P(X_{[i]} = x \mid Y = +1)}{P(X_{[i]} = x, \mid Y = -1))} = d \cdot f_1(\mathbf{X})$$

Overconfident due to cond. Ind. Assumption!

Gaussian Bayes classifiers (GBC)

- Model class label as generated from categorical variable

$$P(Y = y) = p_y, \;\; y \in \mathcal{Y} = \{1, \ldots, c\}$$

- Model features as multivariate Gaussians

$$P(\mathbf{x} \mid y) = \mathcal{N}(\mathbf{x}; \mu_y, \Sigma_y)$$

- Example:
  - Gaussian Naive Bayes (GNB) as special case: $\Sigma_y = \mathrm{diag}\left(\sigma^2_{y,[1]}, \ldots, \sigma^2_{y,[d]}\right)$

- How do we estimate the parameters?

MLE for GBC

- Given data $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- MLE for class label distribution

$$\hat{P}(Y = y) = \hat{p}_y = \frac{\text{Count}(Y = y)}{n}$$

- MLE for feature distribution:

$$\hat{P}(\mathbf{x} \mid y) = \mathcal{N}(\mathbf{x}; \hat{\mu}_y, \hat{\Sigma}_y^2)$$

$$\hat{\mu}_y = \frac{1}{\text{Count}(Y = y)} \sum_{j:y_j = y} \mathbf{x}_j, \quad \hat{\Sigma}y = \frac{1}{\text{Count}(Y = y)} \sum_{j:y_j = y} (\mathbf{x}_j - \hat{\mu}_y)(\mathbf{x}_j - \hat{\mu}_y)^\top$$

Discriminant functions for GBC

- Given $P(Y = +1) = p_+$; $P(\mathbf{x} \mid y) = \mathcal{N}(\mathbf{x}; \mu_y, \Sigma_y)$
- GBC is given by

$$f(\mathbf{x}) = \log \frac{P(Y = +1 \mid \mathbf{x})}{P(Y = -1 \mid \mathbf{x})}$$

$$= \log \frac{p_+}{1 - p_+} + \frac{1}{2} \log \frac{\left|\hat{\Sigma}_-\right|}{\left|\hat{\Sigma}_+\right|} +$$

$$\frac{1}{2} \left[\left((\mathbf{x} - \hat{\mu}_-)^\top \hat{\Sigma}_-^{-1} (\mathbf{x} - \hat{\mu}_-) \right) - \left((\mathbf{x} - \hat{\mu}_+)^\top \hat{\Sigma}_+^{-1} (\mathbf{x} - \hat{\mu}_+) \right) \right]$$

Fisher's linear discriminant analysis (LDA), c = 2

- Suppose we fix $p_+ = 0.5$
- Further, assume covariances are equal: $\Sigma_+ = \Sigma_- = \Sigma$
- Then the discriminant function for GBC could be simplified as

$$
\begin{aligned}
f(\mathbf{x}) &= \log \frac{p_+}{1-p_+} + \frac{1}{2}\left[\log \frac{\left|\hat{\Sigma}_-\right|}{\left|\hat{\Sigma}_+\right|} + \left((\mathbf{x}-\hat{\mu}_-)^\top \hat{\Sigma}_-^{-1} (\mathbf{x}-\hat{\mu}_-)\right) - \left((\mathbf{x}-\hat{\mu}_+)^\top \hat{\Sigma}_+^{-1} (\mathbf{x}-\hat{\mu}_+)\right)\right] \\
&= \frac{1}{2}\left[\left((\mathbf{x}-\hat{\mu}_-)^\top \hat{\Sigma}^{-1} (\mathbf{x}-\hat{\mu}_-)\right) - \left((\mathbf{x}-\hat{\mu}_+)^\top \hat{\Sigma}^{-1} (\mathbf{x}-\hat{\mu}_+)\right)\right] \\
&=
\end{aligned}
$$

Fisher's LDA

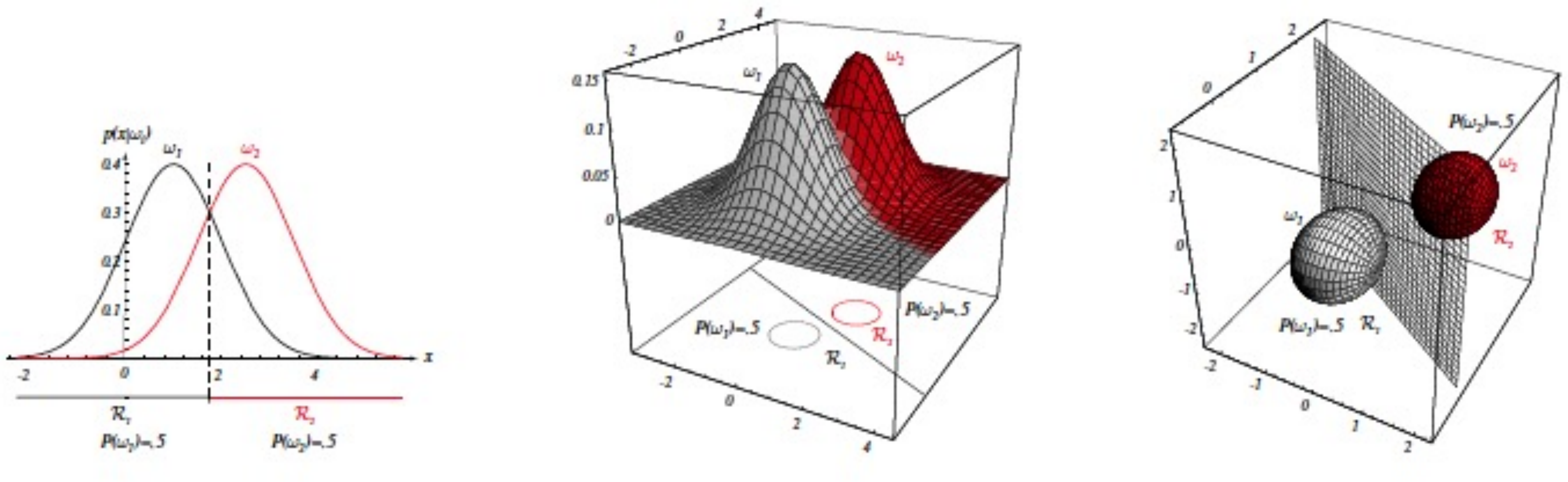
- Assuming
  - binary classification $\mathcal{Y} = \{-1, +1\}$
  - equal class probabilities $p_+ = 0.5$
  - equal covariances $\Sigma_+ = \Sigma_- = \Sigma$

- Fisher's LDA predicts

$$y = \mathrm{sign}(f(\mathbf{x})) = \mathrm{sign}(\mathbf{w}^\top \mathbf{x} + b)$$

where $\mathbf{w} = \hat{\Sigma}^{-1}(\hat{\mu}_+ - \hat{\mu}_-)$ and $b = \frac{1}{2}\left(\hat{\mu}_-^\top \hat{\Sigma}^{-1} \hat{\mu}_- - \hat{\mu}_+^\top \hat{\Sigma}^{-1} \hat{\mu}_+\right)$

LDA Illustration



LDA vs logistic regression

- Fisher's LDA uses the discriminant function

$$f(\mathbf{x}) = \log \frac{P(Y=+1 \mid \mathbf{x})}{P(Y=-1 \mid \mathbf{x})} := \mathbf{w}^\top \mathbf{x} + b$$

$$\Leftrightarrow \quad P(Y=+1 \mid \mathbf{x}) = \frac{1}{1+\exp\left(-f(\mathbf{x})\right)} = \sigma(f(\mathbf{x}))$$

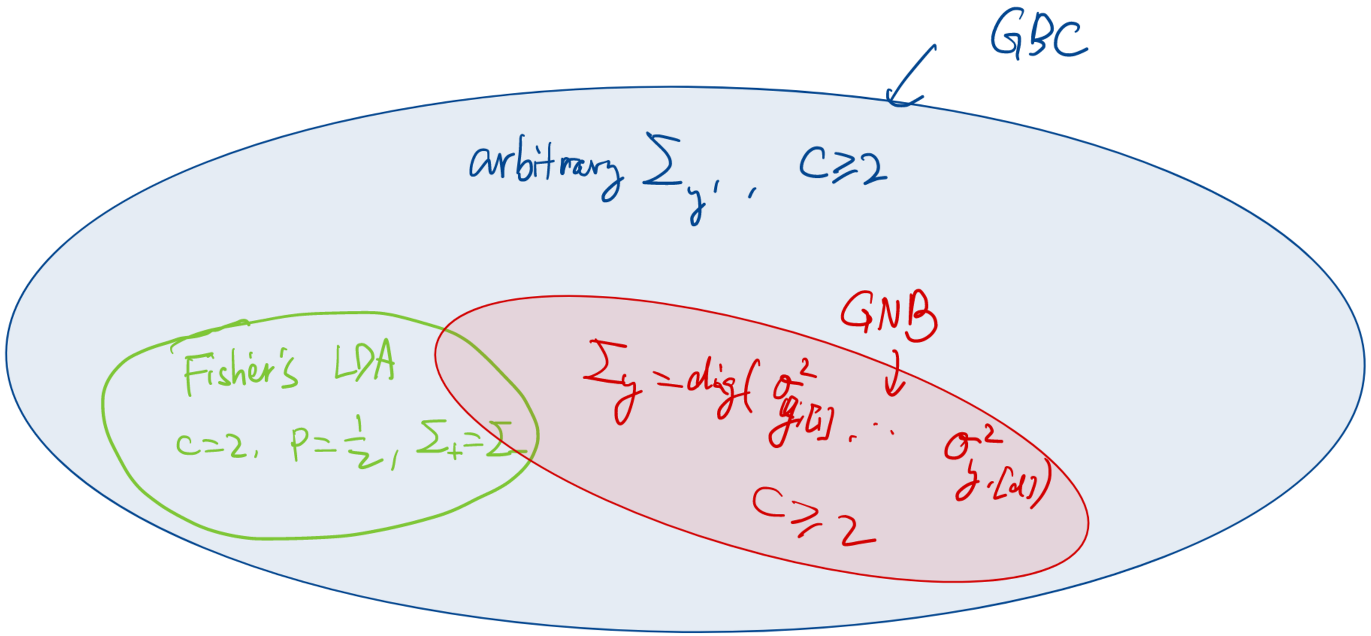
- Therefore, the class probability of LDA is

$$P(Y=+1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

This is of the same form as logistic regression.

- If model assumptions are met, LDA will make same predictions as Logistic Regression!

Gaussian Bayes classifiers



- logistic Reg overlaps with Fisher's LDA (if modeling assumption holds).

  in general, they can give different results.

Fishers LDA vs logistic regression

- Fisher's LDA
  - Generative model, i.e., models P(X,Y)
  - Can be used to detect outliers: P(X) < t
  - Assumes normality of X
  - not very robust against violation of this assumption

- Logistic regression
  - Discriminative model, i.e., models P(Y | X) only
  - Cannot detect outliers
  - Makes no assumptions on X
  - More robust

GNB vs GBC

- Gaussian Naive Bayes models
  - Conditional independence assumption may lead to overconfidence
  - Predictions might still be useful
  - #parameters = $O(cd)$
  - Complexity (memory + inference) linear in d

- General Gaussian Bayes models
  - Captures correlations among features
  - Avoids overconfidence
  - #parameters = $O(cd^2)$
  - Complexity quadratic in d

Avoid overfitting

- Maximum Likelihood Estimation is prone to overfitting
- We can avoid over fitting by
  - Restricting model class, which often leads to fewer parameters
  - Using priors, which often leads to “smaller” parameters

Prior over parameters (c = 2)

- As prior for our class probabilities, have assumed $P(Y = +1) = \theta$
- MLE:

$$\hat{\theta} = \frac{\text{Count}(Y = 1)}{n}$$

  - What happens in the extreme case n = 1?

- May want to put prior distribution P(θ) and compute posterior distribution $P(\theta \mid y_1, \ldots, y_n)$
- Example: Beta prior over parameters

$$\text{Beta}(\theta; \alpha_+, \alpha_-) = \frac{1}{B(\alpha_+, \alpha_-)} \theta^{\alpha_+ - 1} (1 - \theta)^{\alpha_- - 1}$$

Recall: Conjugate distributions

- A pair of prior distributions and likelihood functions is called conjugate if the posterior distribution remains in the same family as the prior.

- Example: Beta priors and Binomial likelihood
  - Prior: $$\mathrm{Beta}(\theta; \alpha_+, \alpha_-)$$
  - Observations: supose we have n_+ positive and n_- negative labels
  - Posterior: $$\mathrm{Beta}(\theta; \alpha_+ + n_+, \alpha_- + n_-)$$
  - Therefore α_+ and α_- act as pseudo-counts. The MAP estimate is

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid y_1, \ldots, y_n; \alpha_+, \alpha_-) = \frac{\alpha_+ + n_+ - 1}{\alpha_+ + n_+ + \alpha_- + n_- - 2}$$

Summary

- Understand connection between discriminative and generative classification
  - Which paradigm is more powerful?
  - Which is in general more robust?

- Relate different Gaussion Bayes classifiers
  - Naïve Bayes
  - Fisher's LDA
  - General GBCs

- Use (conjugate) priors as regularizers
- Compute distributions over features, and use them for outlier detection

Supervised and unsupervised learning summary

Representation/ features	Linear hypotheses, nonlinear hypotheses through feature transformations	
Paradigm	Discriminative vs. generative	
Probabilistic / Optimization Model	Likelihood * Loss function + squared loss = Gaussian lik., 0/1, logistic loss = Bernoulli lik., cross-entropy loss = categorical lik.	Prior Regularization L_2 norm = Gaussian prior, , L_1 norm = Laplace prior, L_0 norm, Beta priors (for Binomial lik)
Method	Exact solution, gradient descent, Bayesian model averaging …	
Evaluation metric	Mean squared error, accuracy, log-likelihood on validation set …	
Model selection	Monte Carlo cross validation, k-fold cross validation, Bayesian model selection …	

References & acknowledgement

- K. Murphy (2021). “Probabilistic Machine Learning: An Introduction”
    - 9.3 “Naive Bayes classifiers”
    - 9.2.1-9.2.4 “Gaussian discriminant analysis”
    - 9.4 “Generative vs discriminative classifiers”

- A. Krause, “Introduction to Machine Learning” (ETH Zurich, 2019)