NAÏVE BAYES

CS 334: Machine Learning

HOMEWORK #3

- Due **10/17 @ 11:59 PM ET** on Gradescope
- 4 questions
  - Feature selection
  - Closed form Linear Regression
  - SGD-based Linear Regression
  - Comparison of closed form and SGD

REMINDER: PROJECT

- Proposal due 10/23: 1-2 pages of problem, dataset, what you plan to do
- Spotlight slides due 10/30
- Spotlight: 11/1 in class
- Presentation: 11/29 and 12/4
- Report and deliverable due 12/13

Samples posted on Piazza

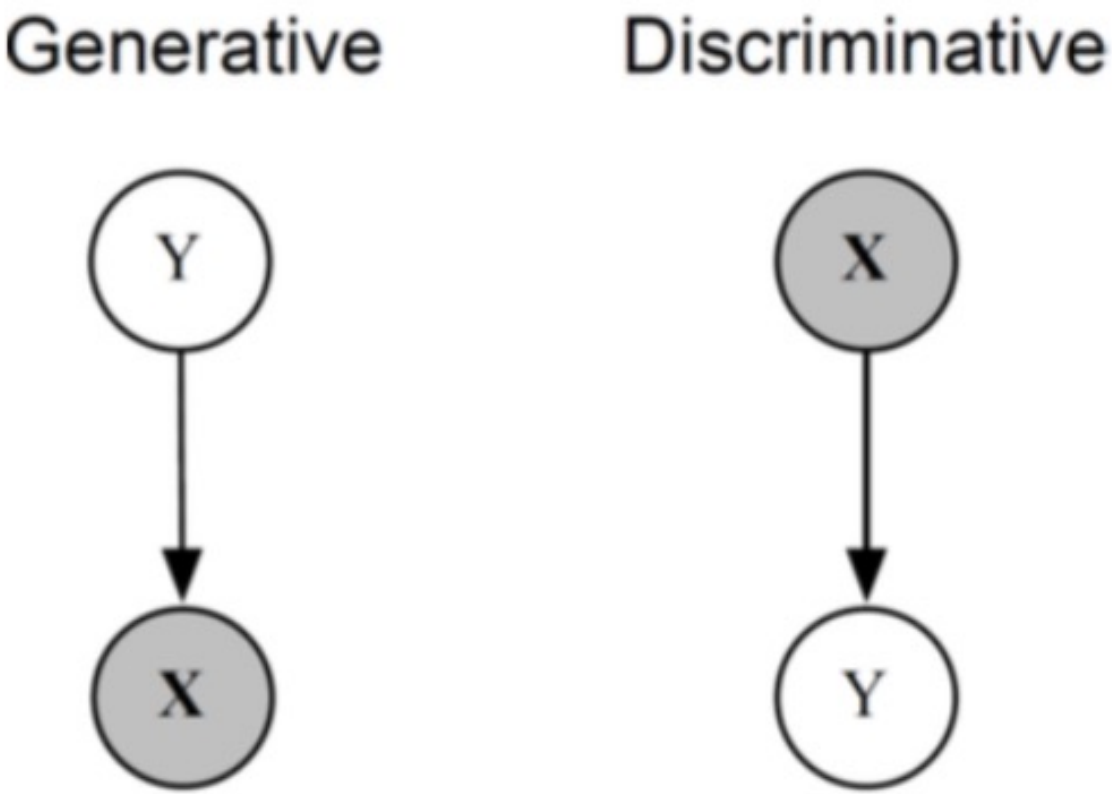
ALTERNATIVE CLASSIFICATION APPROACH

- **Discriminative**: Estimate decision boundary directly from labeled samples
  - Examples: Logistic Regression, Decision Tree
- **Generative**: Model the distribution of input characteristic of the class
  - Examples: Naïve Bayes, Bayesian Networks, Hidden Markov Models

GENERATIVE VS DISCRIMINATIVE ANALOGY

- Task is to determine the language someone is speaking
  - Generative: Learn each language and determine which language the speech belongs to
  - Discriminative: Determine the linguistic differences without learning any language

GENERATIVE VS DISCRIMINATIVE

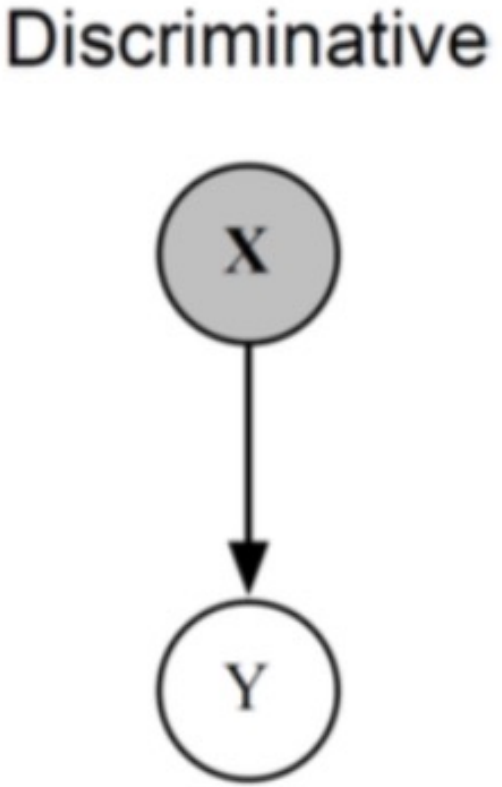


LOGISTIC REGRESSION

- Parameterize the posterior probability

$$P(y = 1|\mathbf{x}, \beta) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}\boldsymbol{\beta})}$$

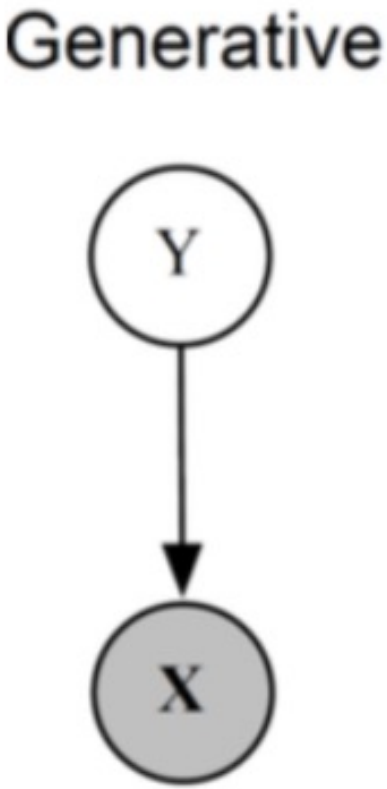
- Learning: learn the parameters (decision boundary)
- Prediction: use the parameters for prediction



NAIVE BAYES

- Learning: model prior P(Y) and likelihood P(X|Y)
- Prediction: Use Bayes rule to calculate the posterior P(Y |X)

$$posterior = \frac{prior \times likelihood}{evidence} \Rightarrow P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)}$$



GROUP ACTIVITY

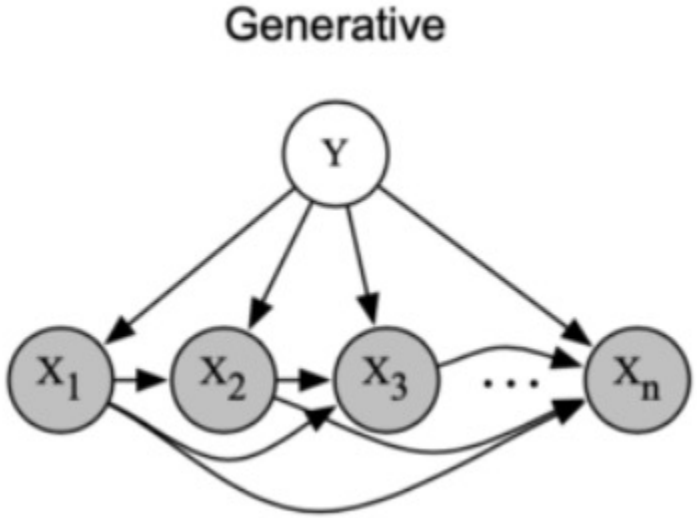
BAYES RULE

$$posterior = \frac{prior \times likelihood}{evidence} \Rightarrow P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)}$$

- Red jar: 10 chocolate + 30 plain
- Yellow jar: 20 chocolate + 20 plain
- Randomly pick a jar, and then randomly pick a cookie
- If it's a plain cookie, what's the probability the cookie is picked out of red jar?

MULTIPLE VARIABLES

- Most machine learning problems contain multiple random variables (RVs)
- Values are not always independent
  - Example: Height (x1) and weight (x2) of newborn
- How to model likelihood P(X|Y)

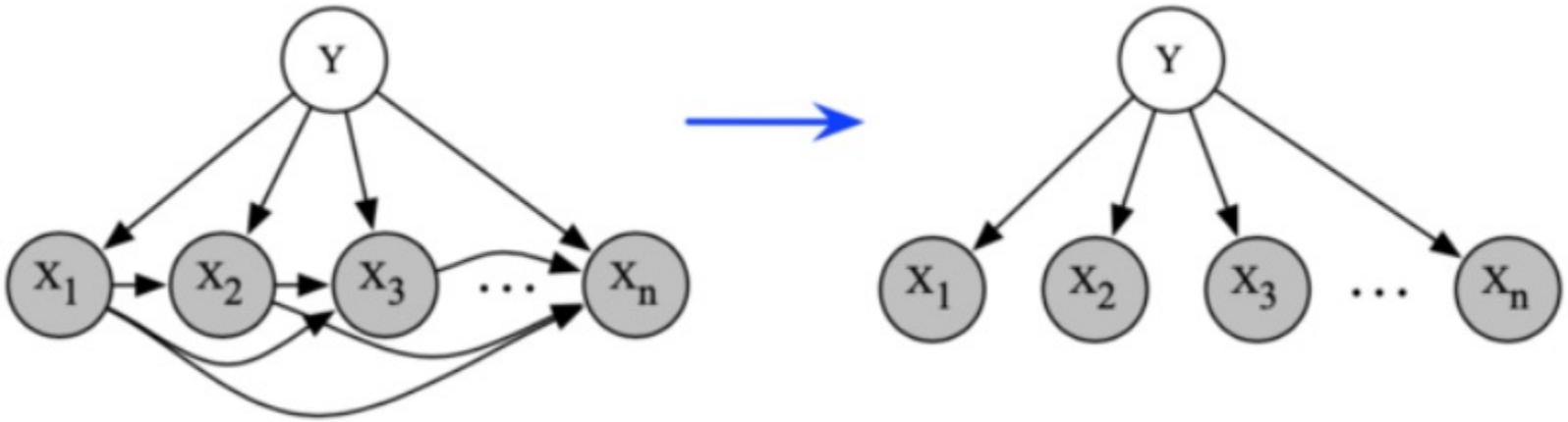


$$posterior = \frac{prior \times likelihood}{evidence} \Rightarrow P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)}$$

NAÏVE BAYES

- Assumes the variables are independent

$$P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$$



NAÏVE BAYES CLASSIFIER: PREDICTION

- Classify using highest a posteriori probability

$$f(\mathbf{x}) = \underset{j=1,\ldots,K}{\arg\max} P(G = k|\mathbf{X} = \mathbf{x})$$

- Application of Bayes' rule:

$$P(G = k|\mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x}|G = k)P(G = k)}{P(\mathbf{X} = \mathbf{x})}$$

- Since denominator same across all classes

$$f(\mathbf{x}) = \underset{j=1,\ldots,K}{\arg\max} P(\mathbf{X} = \mathbf{x}|G = k)\pi_k$$

- Apply independent assumption (naïve Bayes rule)

$$\hat{G}(\mathbf{x}) = \underset{k\in\{1,\cdots,K\}}{\arg\max}\ p(G = k)\prod_{i=1}^{p} p(x_i|G = k)$$

NAÏVE BAYES: DETAILS

- Prediction phase

$$\hat{G}(\mathbf{x}) = \underset{k \in \{1, \cdots, K\}}{\arg\max}\ p(G = k) \prod_{i=1}^{p} p(x_i | G = k)$$

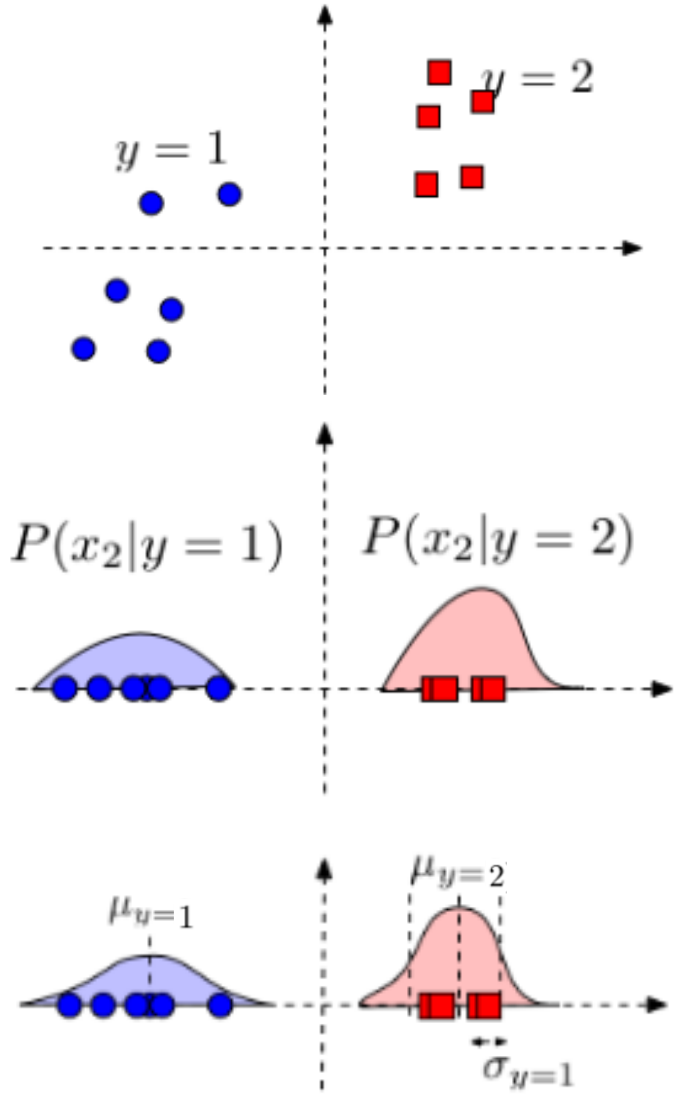
- Learning phase
  - Learn *p(G=k)* and *p(xi|G=k)*

Symptom	Occupation	Ailment
sneeze	nurse	flu
sneeze	farmer	hayfever
headache	builder	concussion
headache	builder	flu
sneeze	teacher	flu
headache	teacher	concussion
sneeze	builder	???

Feature Type	Distribution
Real-valued	Gaussian
Binary	Bernoulli
Categorical	Multinomial

REAL-VALUED FEATURES

- Learn *p(xi|G=k)*
- Use Gaussian (normal) distribution for the feature
- Estimate the mean and variance using data



$$\mu_{\alpha c} \leftarrow \frac{1}{n_c} \sum_{i=1}^{n} I(y_i = c) x_{i\alpha}, \text{where } n_c = \sum_{i=1}^{n} I(y_i = c)$$

$$\sigma^2_{\alpha c} \leftarrow \frac{1}{n_c} \sum_{i=1}^{n} I(y_i = c)(x_{i\alpha} - \mu_{\alpha c})^2$$

CATEGORICAL FEATURES

Symptom	Occupation	Ailment
sneeze	nurse	flu
sneeze	farmer	hayfever
headache	builder	concussion
headache	builder	flu
sneeze	teacher	flu
headache	teacher	concussion
sneeze	builder	???

$$\hat{G}(\mathbf{x}) = \underset{k \in \{1, \cdots, K\}}{\arg\max}\ p(G = k) \prod_{i=1}^{p} p(x_i | G = k)$$

- Learn $p(xi|G=k)$
- Count the frequency of each category given a class

$$\frac{\#\text{ of samples with label c that have feature } \alpha \text{ with value } j}{\#\text{ of samples with label } c}$$

- What if some attribute values do not appear in training data?

GROUP ACTIVITY

EXAMPLE: CATEGORICAL FEATURES

Symptom	Occupation	Ailment
sneeze	nurse	flu
sneeze	farmer	hayfever
headache	builder	concussion
headache	builder	flu
sneeze	teacher	flu
headache	teacher	concussion
sneeze	builder	???

Predict the ailment of the new record using naïve bayes

$$\hat{G}(\mathbf{x}) = \underset{k \in \{1, \cdots, K\}}{\arg\max}\ p(G = k) \prod_{i=1}^{p} p(x_i | G = k)$$

NAÏVE BAYES: PRACTICAL ISSUES

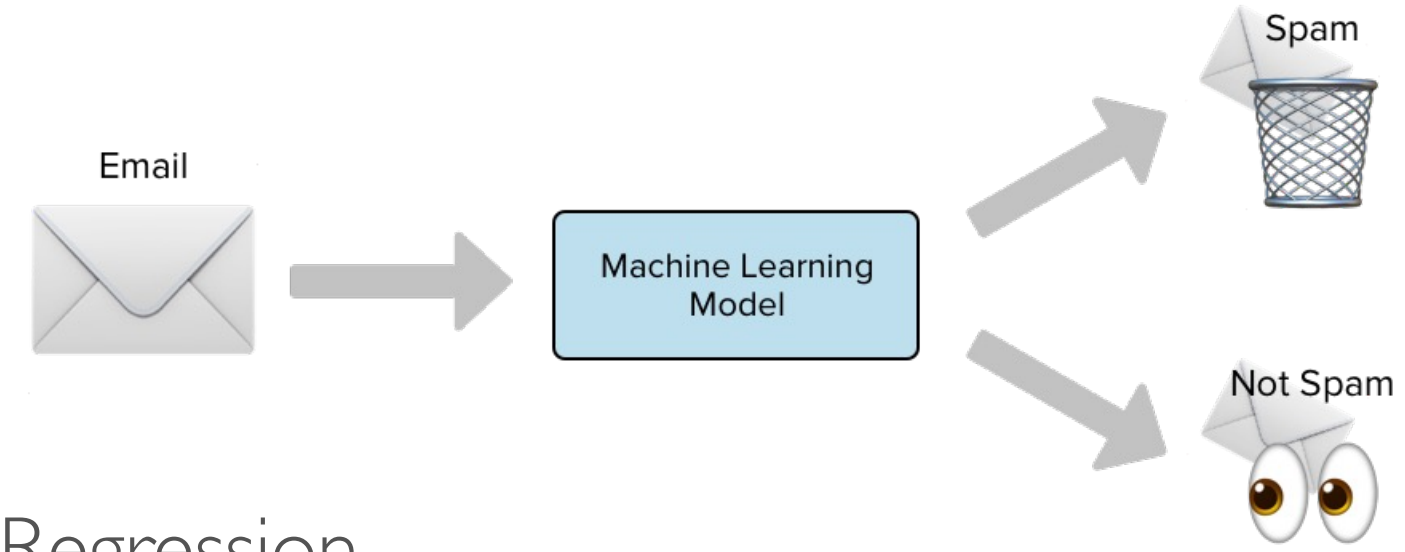
- Multiplied likelihood probabilities become too small
  - Instead of multiplying likelihood probabilities, use log likelihood and add them
- Some attribute values may not appear in training data
  - Smoothing: add a small probability value for each attribute value – like adding pseudo-observations to the training data

NAÏVE BAYES: SUMMARY

- Advantages
  - Fast to train (single scan) and classify
  - Not sensitive to irrelevant features
  - Handles all types of data well
- Disadvantages
  - Assumes feature independence

HOMEWORK #4

- Out 10/13 and due **10/27 @ 11:59 PM ET** on Gradescope
- 3 questions
  - Feature extraction
  - Perceptron
  - Naïve Bayes and Logistic Regression

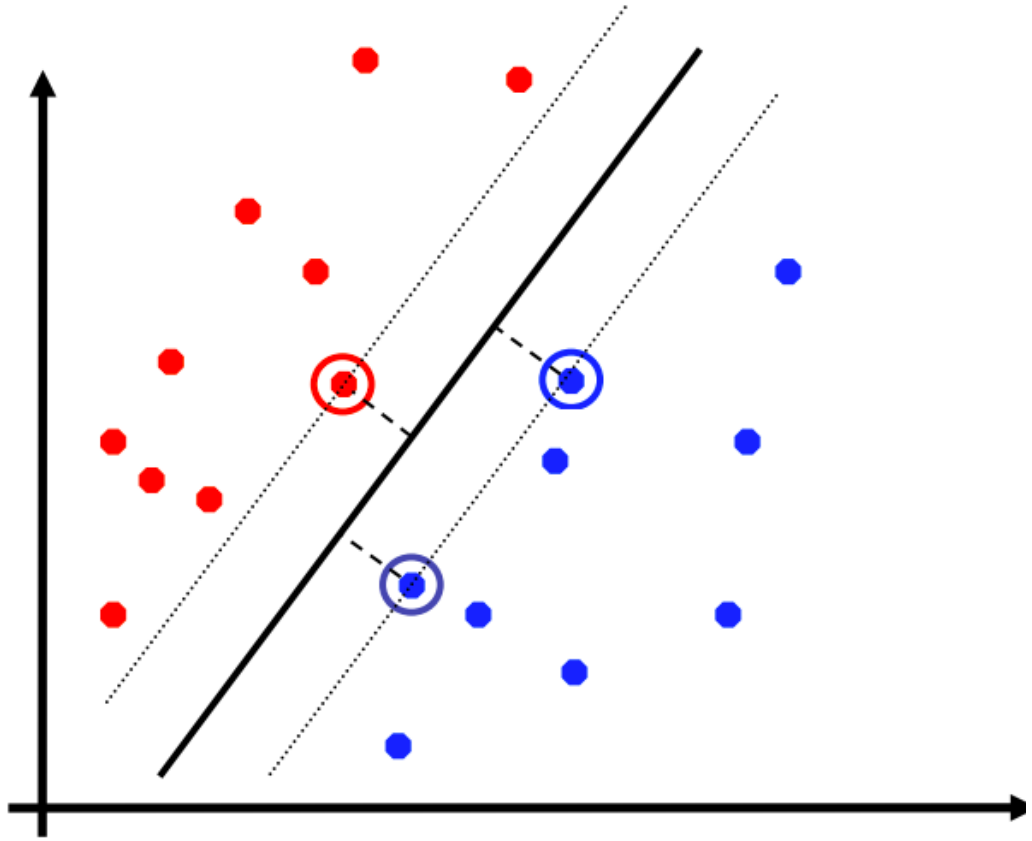


SUPPORT VECTOR MACHINES

CS 334: Machine Learning

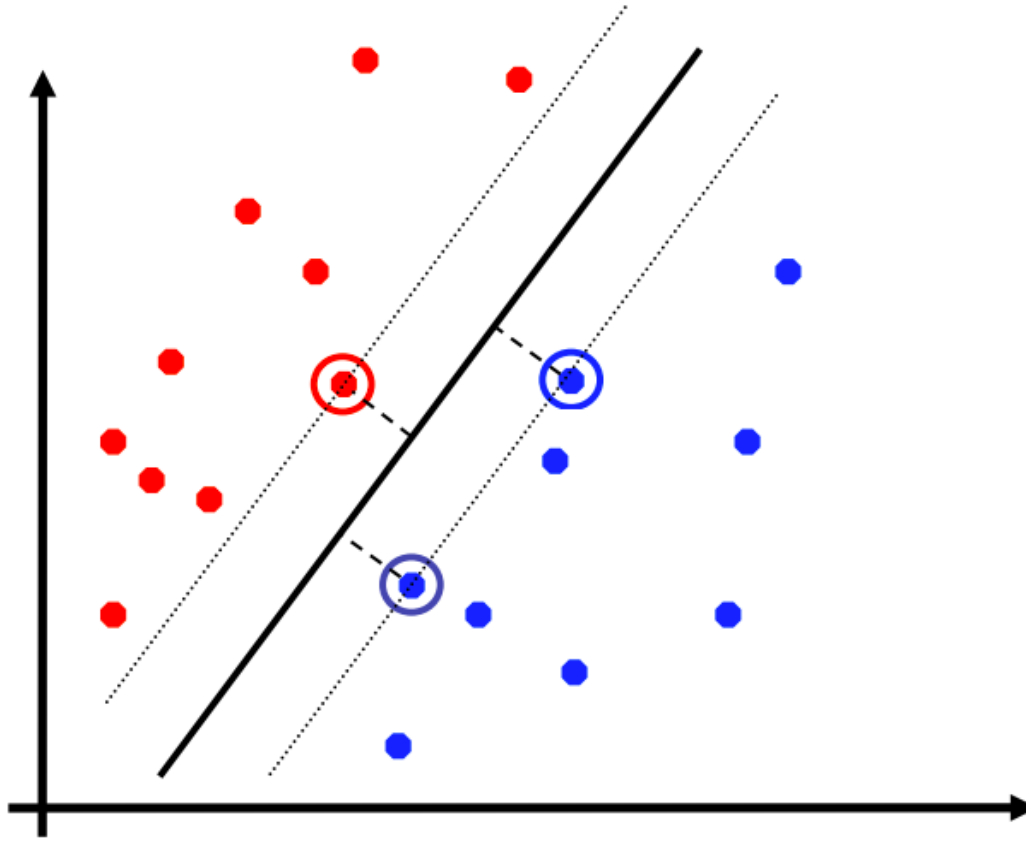
SUPPORT VECTOR MACHINE (SVM)

- Introduced by Boser, Guyon, and Vapnik in 1992
- Chose the linear separator with the largest margin
  - Robust to outliers
  - Good according to intuition, theory, and practice



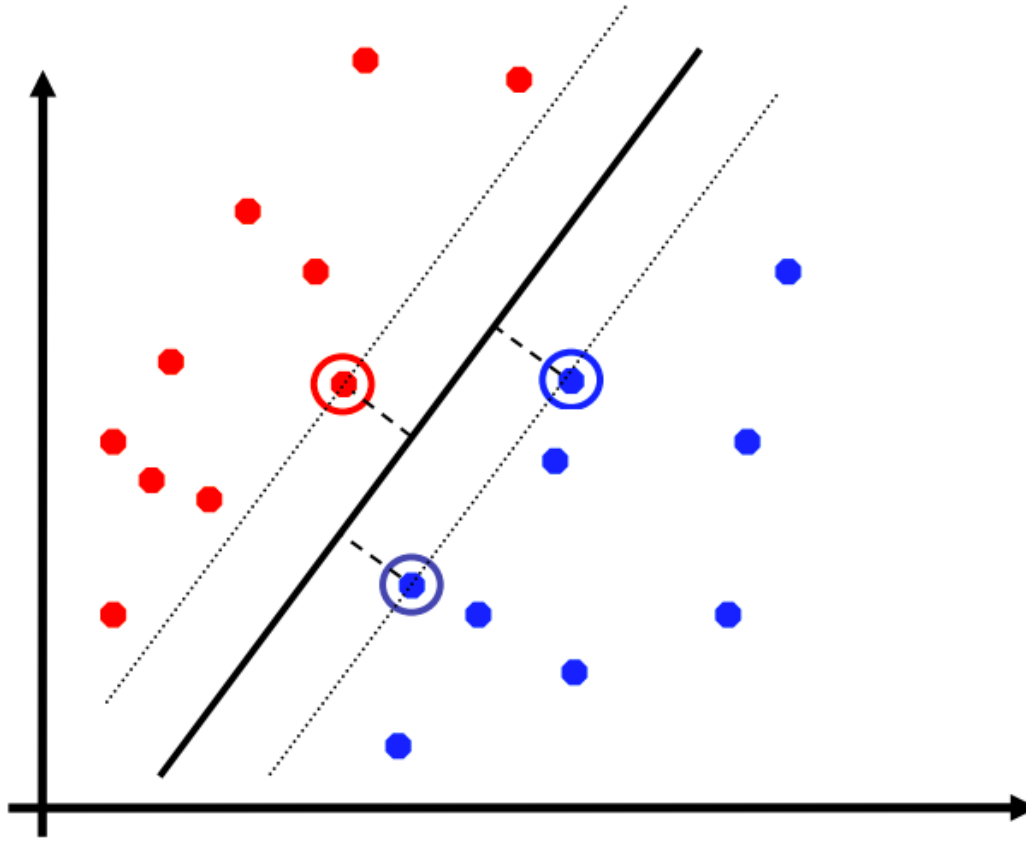
SVM: KEY IDEAS

- Find large margin separator to improve generalization
- Use soft margin and optimization to find solution with few errors
- Use kernels to make large feature spaces computationally efficient



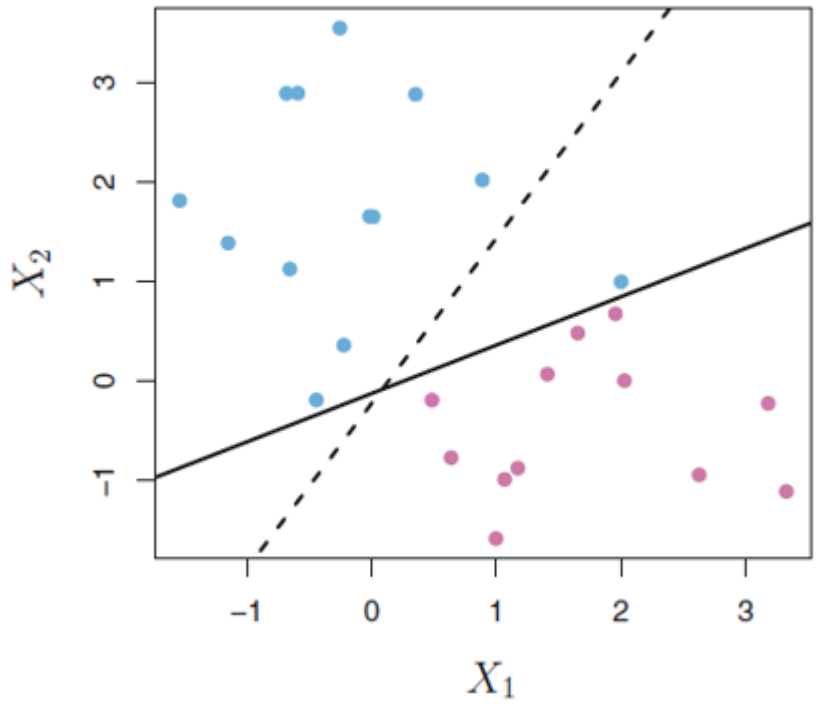
SVM KEY IDEA #1: MAXIMAL MARGIN

- Find the hyperplane that is farthest from the training observations (maximal margin)
- Margin is the minimal distance from the observations to the hyperplane



SOFT MARGIN CLASSIFIER

- Allow some observations to be on the incorrect side of the margin or hyperplane (soft margin)
- Greater robustness to individual observations
- Better classification of *most* of the training observations

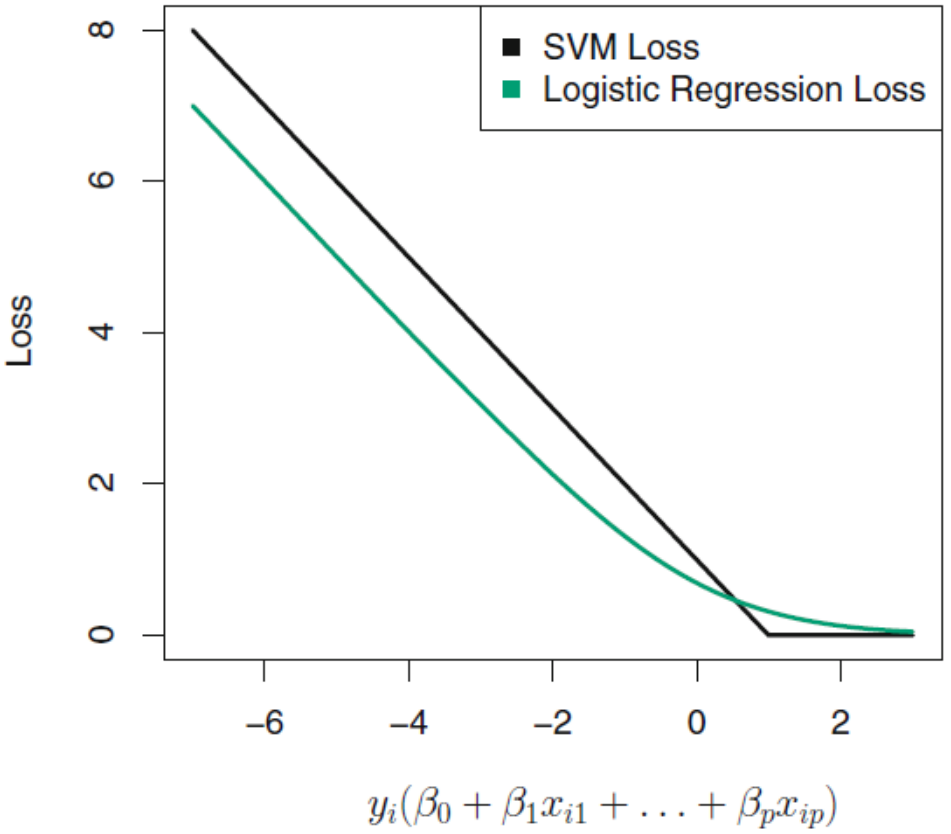


SOFT MARGIN CLASSIFIER

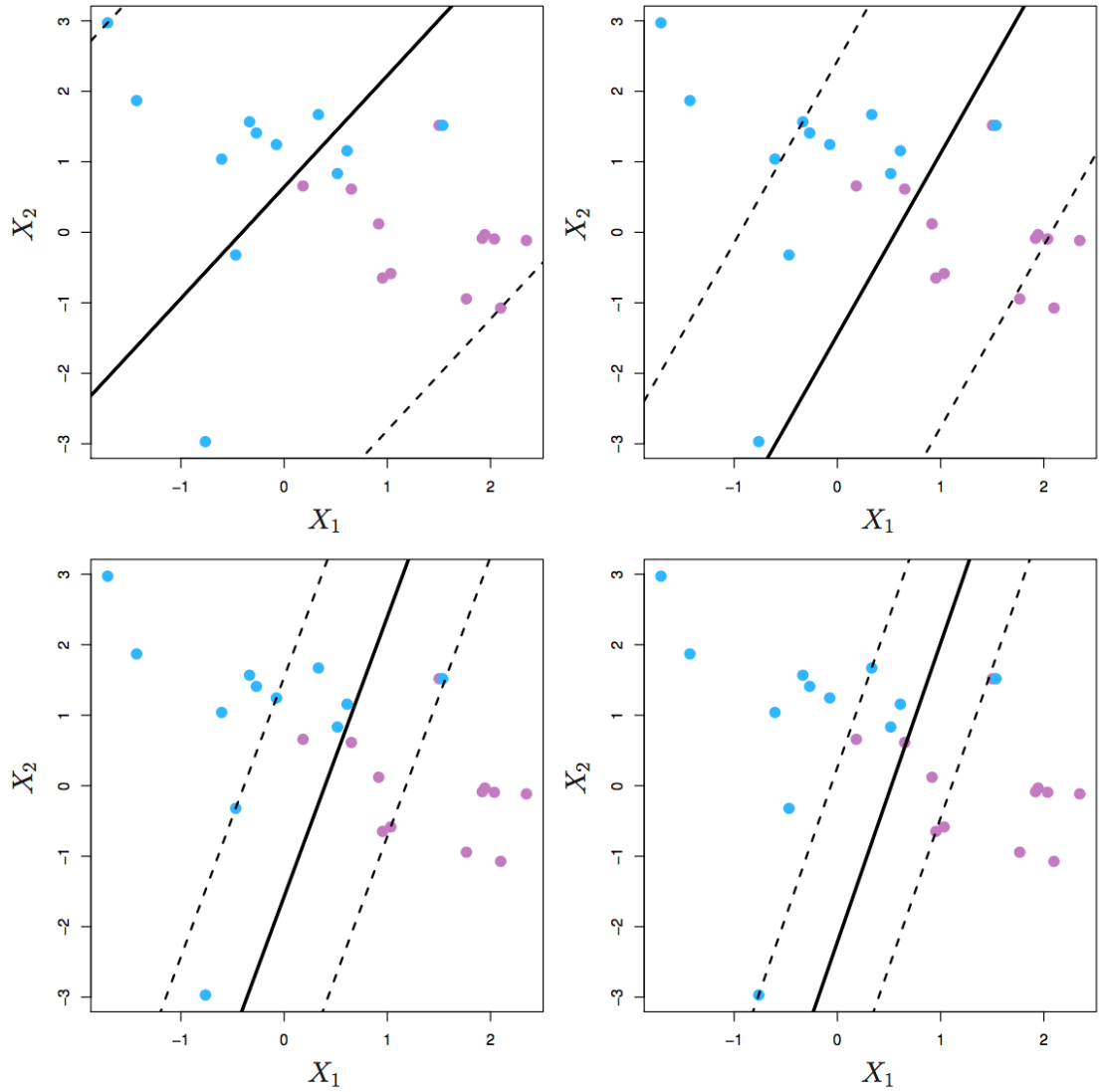
- SVM unconstrained optimization problem

$$\min \left(||\boldsymbol{\beta}||_2^2 + C \sum_i \max(0, 1 - y_i(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}_i)) \right)$$

- regularization term + hinge loss
- The choice of tuning parameter C is very important



C REGULARIZATION PARAMETER



As C increases, margin decreases, bias decreases, variance increases

KEY IDEA #3: KERNELS

- Solve for hyperplane in high dimensional space where data is separable
- If D is very large, many more parameters to learn than in original space. Can we use just the data points to learn the separating hyperplane?

KEY IDEA #3: KERNELS

- Solution: $f(\mathbf{x}) = \sum_j \alpha_j y_j \, \mathbf{x}_j \cdot \mathbf{x} + \beta_0$
- Solution with feature mapping:

$$f(\mathbf{x}) = \sum_j \alpha_j y_j \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}) + \beta_0$$

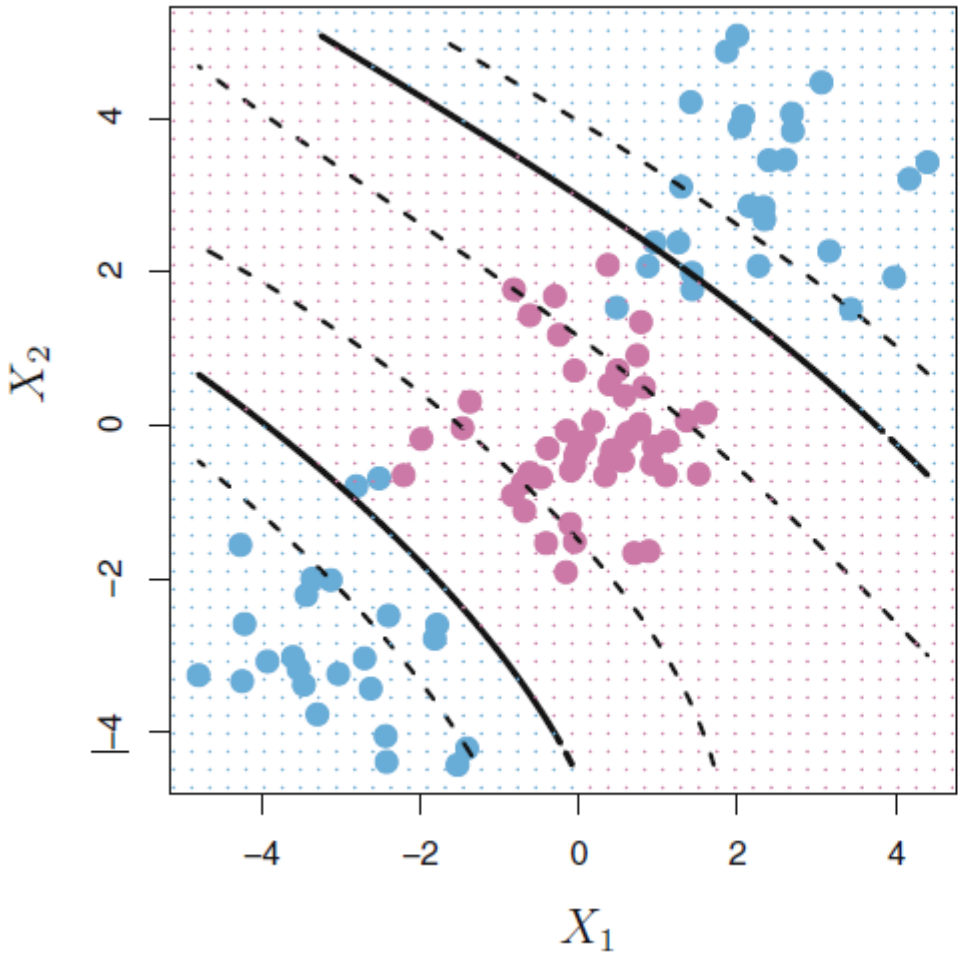
- A *kernel function* corresponds to a dot product of two feature vectors in some expanded feature space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

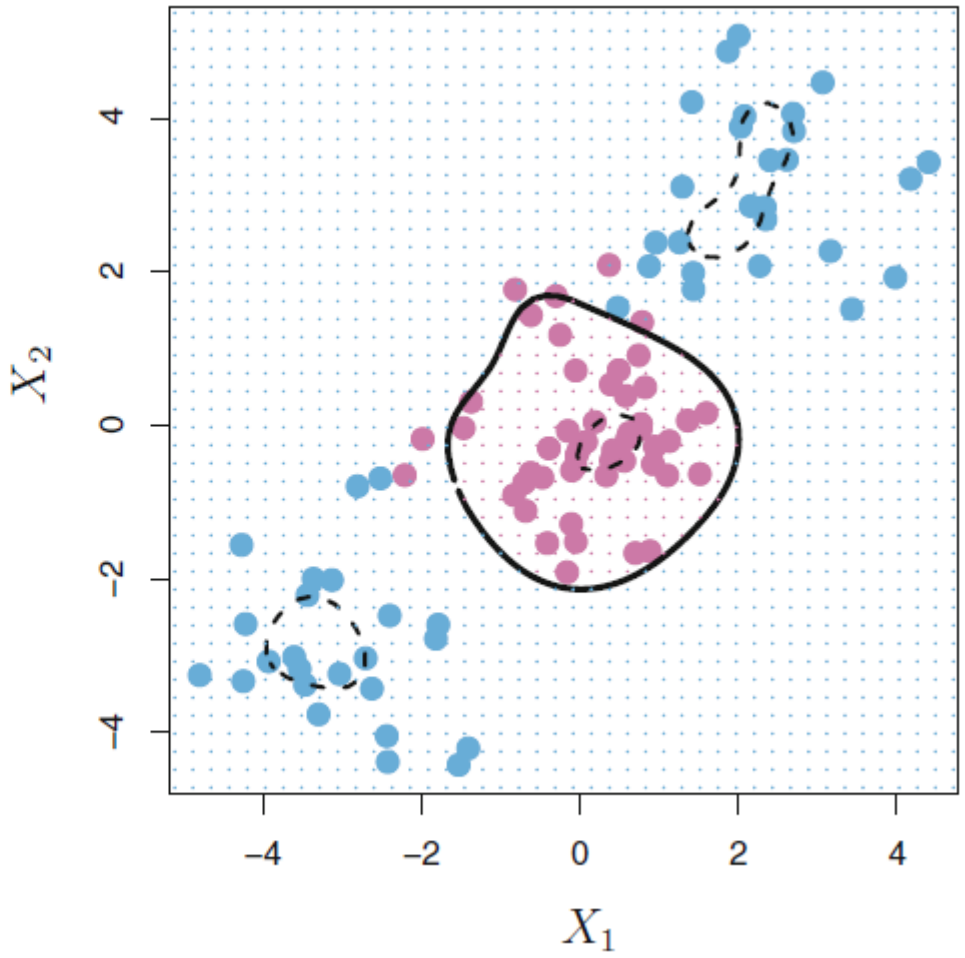
- Think of kernel measure as similarity

KERNELS

Polynomial d=3



Radial



BENEFITS OF KERNELS

- Efficient: often times easier than computing feature map and then dot product
  - Especially from memory perspective — need to store less
- Flexibility: function chosen arbitrary so long as existence of feature map is guaranteed