Naïve Bayes

Background

There are three methodologies:

a) Model a classification rule directly

Examples: k-NN, SVM, neural nets, ..

b) Model the probability of class memberships given input data

Examples: logistic regression, probabilistic neural nets (softmax),…

c) Make a probabilistic model of data within each class

Examples: naive Bayes, GMM

Important ML taxonomy for learning models

probabilistic vs non-probabilistic models

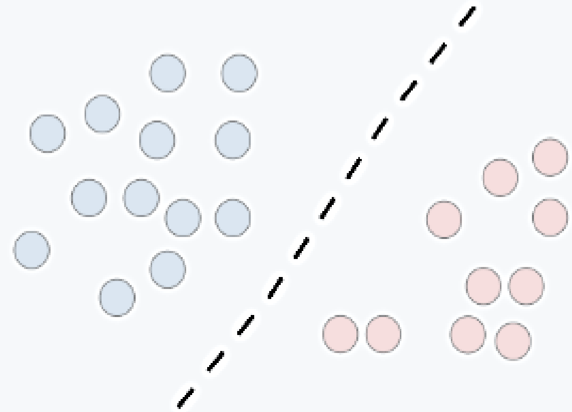
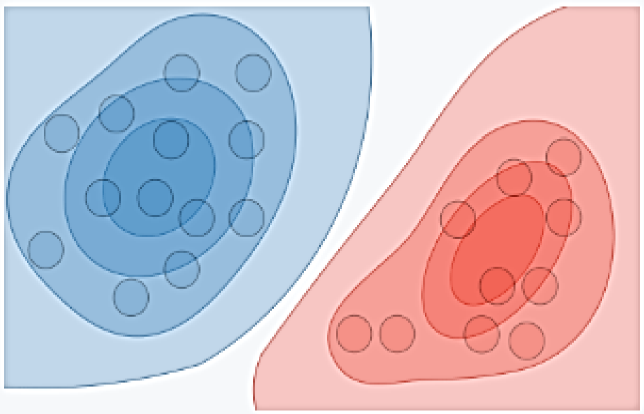
discriminative vs generative models

Background

Based on the taxonomy, we can see different essence of learning models (classifiers) more clearly.

ML Taxonomy	Probabilistic	Non-Probabilistic
Discriminative	• Logistic Regression • Probabilistic neural nets • ……..	• kNN • Linear classifier • SVM • ……
Generative	• Naïve Bayes, GMM etc.	(?)

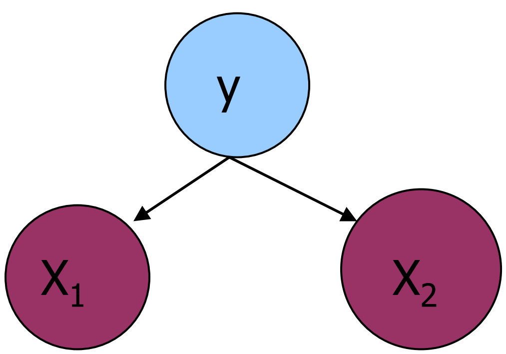
Background

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

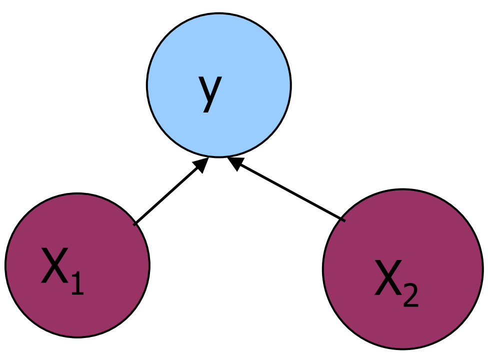
https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning

Comparison

- Generative models
  - Assume some functional form for P(X|y), P(y)
  - Estimate parameters of P(X|y), P(y) directly from training data
  - Use Bayes rule to calculate P(y|X)
- Discriminative models
  - Directly assume some functional form for P(y|X)
  - Estimate parameters of P(y|X) directly from training data



Naïve Bayes
Generative

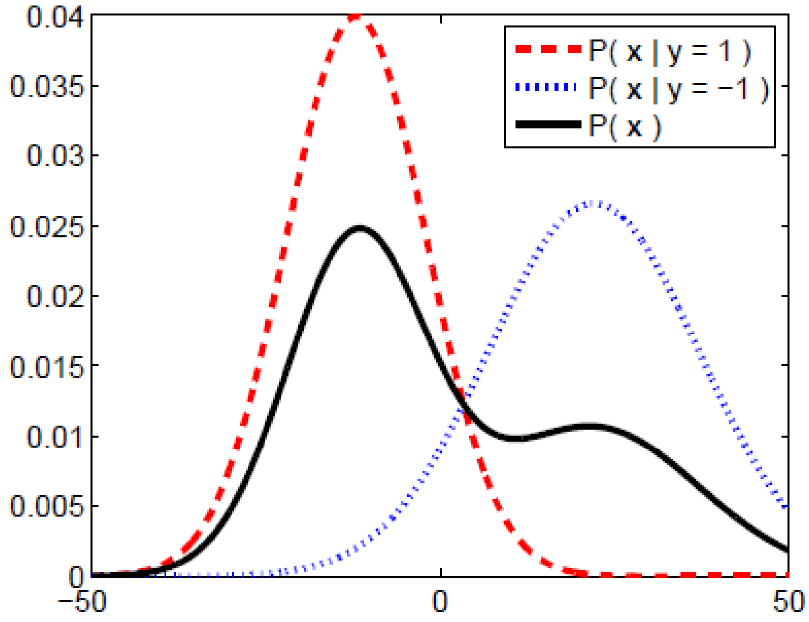


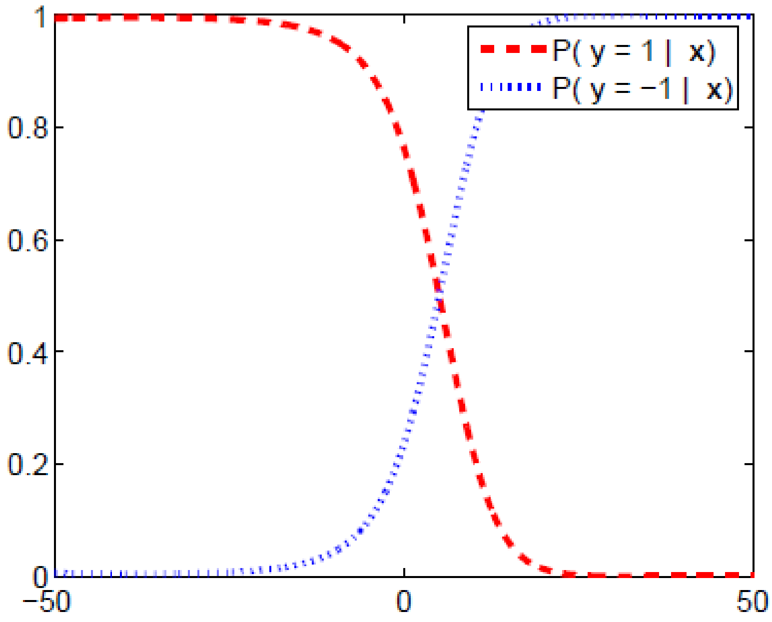
Logistic Regression
Discriminative

Bayes Formula

Bayes, Thomas (1736) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, London, 53:370-418.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

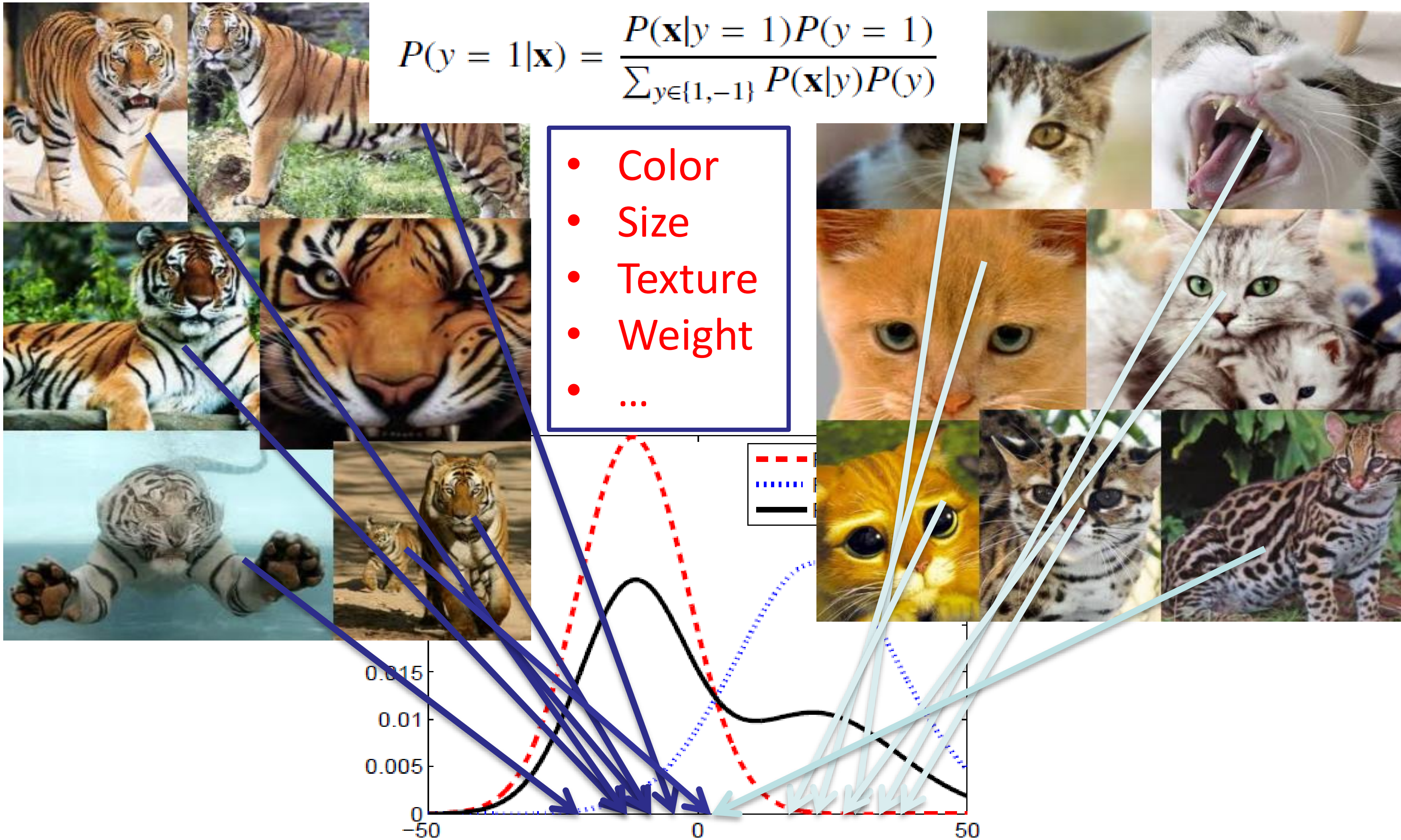




Generative Model

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1)P(y=1)}{\sum_{y\in\{1,-1\}} P(\mathbf{x}|y)P(y)}$$

- Color
- Size
- Texture
- Weight
- …

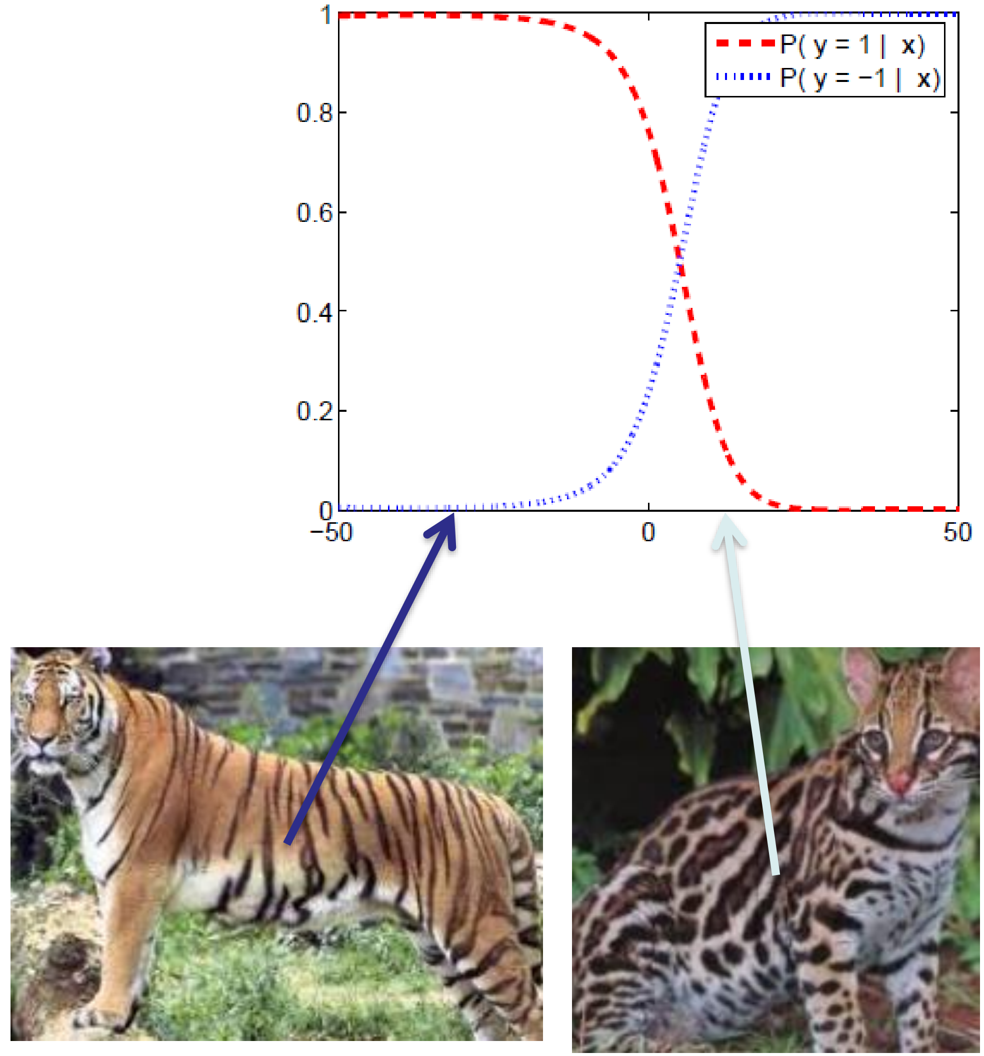


Discriminative Model

Logistic Regression

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

- Color
- Size
- Texture
- Weight
- …



Probability Basics

Prior, conditional and joint probability for random variables

- Prior probability: $P(X)$
- Conditional probability: $P(X_1 \mid X_2), P(X_2 \mid X_1)$
- Joint probability:

  Relationship: $X = (X_1, X_2), P(X) = P(X_1, X_2)$

  $P(X_1, X_2) = P(X_1 | X_2) * P(X_2)$

  Independence: $P(X_1 | X_2) = P(X_1)$,

  $\boldsymbol{P(X_1, X_2) = P(X_1) * P(X_2)}$

$$P(y|\mathbf{X}) = \frac{P(\mathbf{X}|y)P(y)}{P(\mathbf{X})}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Prior is what you believe about some quantity at particular point in time, and the **posterior** is your belief once additional information comes in.

More specifically, the **prior** tells you the relative likelihood of different values of some quantity (a parameter) in the absence of data

The **Posterior** tells you how you'd revise those beliefs "in the presence of data"

posterior probability
(physican's diagnosis)

likelihood function
(medical knowledge)

prior probabilty
(flu season?)

$$P(\text{have the flu} \mid \text{symptoms}) = \frac{P(\text{symptoms} \mid \text{have the flu}) \times P(\text{have the flu})}{P(\text{symptoms})}$$

evidence
(physican exam)

Bayes' Theorem in the Doctor's Office

Bayes Theorem Intuition

Example (Cancer diagnosis)

Initial belief + new evidence -> new and improved belief

$p(C|T) = p(T|C)*p(C) / p(T)$

C for cancer(event) and T for test (evidence)

- p(C|T) - probability of Cancer if Test is positive (***objective***)
- p(T|C) - probability that Test is positive if Cancer is true
- p(C) - probability that Cancer is true
- p(T) - probability that Test is positive

$$p(T) = (p(C)*p(T|C)) + (p(\text{not } C)*p(T|\text{not } C))$$

Example (Cancer diagnosis)

Incidence in the population = 0.01.
Test quality – 0.99 (99% with cancer will test positive, 99% without cancer will test negative)

p(T) - the probability of testing positive *whether or not you have cancer (FP + TP)*.

TP = (pct of people w cancer)*(rate of true positives)
= p(C) * p(T|C) = 0.01*0.99=**0.0099**.
FP = (pct of people w/o cancer)*(rate of false positives)
= 0.99*0.01 = 0.0099

p(T) = TP+FP = 0.0099+0.0099 = **0.0198**

p(C/T) – prob of cancer if test is + is =0.0099/0.0198 = 0.50

(If test is 100% reliable, we don’t need Bayes theorem)

Test Accuracy
0.99

Incidence
0.01

		Pred		
		No	Yes	
Actual	No	9801	99	9900
	Yes	1	99	100
		9802	198	10000

Test Accuracy
0.99

Incidence
0.01

		Pred		
		No	Yes	
Actual	No	0.9801	0.0099	0.99
	Yes	0.0001	**0.0099**	0.01
		0.9802	**0.0198**	1

Why is “Naive Bayes" naive?

A naive Bayes classifier assumes that the presence (or absence) of a particular feature is unrelated to the presence (or absence) of any other feature.

Since this assumption (the absolute independence of features) is probably never met in practice, it is "naive".

$$p(\mathbf{x} \mid \omega_j) = p({}_1 \mid \omega_j) \cdot \ldots \cdot p({}_d \mid \omega_j) = \prod_{k=1}^{d} p(x_k \mid \omega_j)$$

If you like **Pickles**, and you like **Ice Cream**, naive bayes will assume independence and give you a **Pickled Ice Cream** and think that you'll like it.

sklearn Naïve Bayes Algorithms

Gaussian NB: for features in continuous form. GNB assumes features to follow a normal distribution.

MultiNomial NB: for features with discrete values like word count 1,2,3...

Bernoulli NB: for features with binary or boolean values like True/False or 0/1.

Advantages of Naïve Bayes

- A small amount of training data to estimate parameters (means and variances of the variable) - not the entire covariance matrix

- Test is straightforward - calculating conditional probabilities with normal distribution

- the only task before prediction is finding the parameters for the features' individual probability distributions, which can be done quickly and deterministically. This means classifier can perform well with high-dimensional data and/or large data.

- It perform well in case of ***categorical input variables*** compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Disadvantages

- Naive Bayes is also known as a bad estimator, so the probability outputs are not to be taken too seriously.

- Another limitation is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Conclusion

- Performance competitive to most of state-of-art classifiers even in presence of violating independence assumption.
- Many successful application, e.g., spam mail filtering
- A good candidate as a base learner in ensemble learning

```
from sklearn.naive_bayes import BernoulliNB

# instantiate bernoulli NB object
bnb = BernoulliNB()

# fit
bnb.fit(X_train_transformed, y_train)

# predict class
y_pred_class = bnb.predict(X_test_tranformed)

# predict probability
y_pred_proba =bnb.predict_proba(X_test_tranformed)

# accuracy
from sklearn import metrics
metrics.accuracy_score(y_test, y_pred_class)
```

Thank You