Introduction to Classification

CSCI 347

Adiesha Liyana Ralalage

Classification

Input data matrix

Class labels are always categorical

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Goal is to predict the class of the new data

Classification

Input is also commonly in the form:

Weather	Weekend?	Finished HW	Target/Label/Class
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New data instance

Weather	Weekend?	Finished HW	Target/Label /Class
Sunny	Yes	No	?

Classification

Input is also commonly in the form:

Weather	Weekend?	Finished HW
Snow	Yes	No
Overcast	No	No
Sunny	Yes	No
Overcast	Yes	Yes
Overcast	No	Yes
Snow	No	Yes
Overcast	Yes	No
Sunny	Yes	No
Sunny	No	Yes
Snow	No	Yes
Snow	Yes	No
Overcast	Yes	No
Overcast	No	Yes

y
Yes
No
Yes
Yes
Yes
No
No
No
Yes
Yes
No
No
Yes

New data instance

Weather	Weekend?	Finished HW	y
Sunny	Yes	No	y=?

MONTANA STATE UNIVERSITY

Introduction to Bayes classifier

CSCI 347

Bayes Classifier

- You are given a training datatset D of n points.
  - $D = \{x_1, x_2, \dots, x_n\}$
- Each point $x_i \in D$ is in d dimensional space.
- You are also given class labels of each data point x_i, denoted as y_i.
  - $y = \{y_1, y_2, \dots, y_n\}$ *this is the label column*
  - $y_i \in \{c_1, c_2, \dots, c_k\}$ *each label is from one of these k valeus*
- The objective of the Bayes classifier is, given a new data point calculate the $p(c_i|x)$

Let's look at some basic probability theory

- Conditional probability:
  - conditional probability is a measure of the probability of an event occurring, given that another event is already known to have occurred.
  - If we are given two events A and B, then $\mathrm{P}(A|B)$ is the probability of event A occurring given that the event B has occurred.
  - P of A given B
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$

MONTANA STATE UNIVERSITY

Bayes Theorem

- Bayes theorem gives a mathematical rule for inverting conditional probabilities, allowing you to find the probability of a cause given effect.

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

MONTANA STATE UNIVERSITY

Bayes Theorem

- In Bayes classifier, we try to calculate the probability of a class of a new data instance, using the test dataset.
  - $P(c_i|x)?$, where x is the new data point
- Let's try to apply bayes theorem for this probability.
  - $P(c_i|x) = \frac{P(x \mid c_i)P(c_i)}{P(x)}$
  - $P(c_i|x)$ *is called the posterior probability*
  - $P(c_i)$ is called the prior probability
  - $P(x \mid c_i)$ is the likelihood
  - $P(x)$ is the probability of observing x from any of the k classes
  - $P(x) = \sum_{j=1}^{k} P(x|c_j) \cdot P(c_j)$

Bayes Theorem

- Let's try to apply bayes theorem for this probability.
  - $P(c_i|x) = \frac{P(x \mid c_i)P(c_i)}{P(x)}$
  - $P(c_i|x)$ *is called the posterior probability*
  - $P(c_i)$ is called the prior probability
  - $P(x \mid c_i)$ is the likelihood
  - $P(x)$ is the probability of observing x from any of the k classes
  - $P(x) = \sum_{j=1}^{k} P(x|c_j) \cdot P(c_i)$
- We calculate this probability for all classes and pick the class that maximizes the probability.
- $\forall i \in [k]$: $P(c_i|x)$ must be calculated and pick the c_i that maximize the probability.
- $\hat{y} = argmax_{c_i}\{P(c_i|x)\} = argmax_{c_i}\left\{\frac{P(x \mid c_i)P(c_i)}{P(x)}\right\} = argmax_{c_i}\{P(x \mid c_i)P(c_i)\}$

Bayes Theorem

$$\hat{y} = argmax_{c_i}\{P(x \mid c_i)P(c_i)\}$$
$$P(c_i|x) = P(x \mid c_i)P(c_i)$$

- Now I need to calculate these probabilities.
- It is actually very difficult to calculate these probabilities, therefore we could only estimate them.

$$\hat{P}(c_i|x) = \hat{P}(x|c_i)\hat{P}(c_i)$$

- Estimating prior probability $P(c_i)$
  - Let $D_i = \{x_j \in D \mid x_j \ has\ class\ y_j = c_i\}$
  - $|D| = n$, and $|D_i| = n_i$
  - $\hat{P}(c_i) = \frac{|D_i|}{|D|} = \frac{n_i}{n}$

Bayes Theorem

- Estimating prior probability $P(c_i)$
  - Let $D_i = \{x_j \in D \mid x_j \; has \; class \; y_j = c_i\}$
  - $|D| = n$, and $|D_i| = n_i$
  - $\hat{P}(c_i) = \frac{|D_i|}{|D|} = \frac{n_i}{n}$
- There are two ways you can use to estimate the likelihood $P(x|c_i)$
  - Parametric approach
    - We make an assumption about the distribution of a particular class.
    - Ex: Each class is normally distributed.
  - Non-parametric approach
    - I will not go into details about this in this lecture

Bayes Theorem

- There are two ways you can use to estimate the likelihood $P(x|c_i)$
  - Parametric approach
    - We make an assumption about the distribution of a particular class.
    - Ex: Each class is normally distributed.
  - Non-parametric approach
    - I will not go into details about this in this lecture
- Suppose we assume that each class c_i is normally distributed around some mean μ_i with corresponding covariance matrix Σ_i, both of which are estimated by the D_i.
  - $f_i(x) = f(x|\mu_i, \Sigma_i) = \frac{1}{\left(\sqrt{2\pi}\right)^d \sqrt{|\Sigma_i|}} \exp\left\{-\frac{(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)}{2}\right\}$

Bayes Theorem

- Suppose we assume that each class c_i is normally distributed around some mean μ_i with corresponding covariance matrix Σ_i, both of which are estimated by the D_i.
  - $f_i(x) = f(x|\mu_i, \Sigma_i) = \frac{1}{\left(\sqrt{2\pi}\right)^d \sqrt{|\Sigma_i|}} \exp\left\{-\frac{(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)}{2}\right\}$
- Remember that to calculate $f_i(x)$ (which is the estimate of the $P(x|c_i)$), we use D_i.
- But there are issues in this method:
  - When the number of dimensions are very high, we cannot reliably estimate Σ_i (covariance matrix)
  - If you only have few data points for a particular class, then the estimates are not of good quality.
  - In fact, it will appear that the new point comes from all the classes with equal likelihood.
  - Moreover, if you have a class with large number of datapoint, in most cases that class will be chosen as the predicted value.
  - $\mathrm{P}(c_i|x_j) \approx P(c_i)\ which\ means\ your\ data\ has\ no\ influence$
  - For lower dimensional data, this might work even with small number of data points but would perform worse when number of dimensions are high.

Bayes Theorem

- Bayes classifier is the optimal classifier, if you know the probability distribution of all your data.
- But as we discussed earlier, this is difficult to calculate.

Bayes Theorem

- What is the solution?
- We use an approach called naïve bayes.
- This is even simpler method.
- Surprisingly, this works very well for real world data.
- Assumption:
  - We assume that our attributes are all independent

Bayes Theorem

- Assumption:
  - We assume that our attributes are all independent.

$$P(x \mid c_i) = P(x_1, x_2, \dots, x_d \mid c_i) = \prod_{j=1}^{d} P(x_j \mid c_i)$$

Introduction to Naïve Bayes Algorithm

CSCI 347

Naïve Bayes is a classification algorithm

Input data matrix

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

MONTANA STATE UNIVERSITY

Naïve Bayes

Input data matrix

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

Naïve Bayes

Input data matrix

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes

What is the probability of "Yes" and what is the probability of "No" in our data? (These are prior probabilities)

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

We want to estimate the probabilities of c_1 and c_2 **after observing the new data instance** and then **choose the one with the maximum probability.**

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes

What is the probability of "Yes" and what is the probability of "No" **after observing the new data instance?**

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_1|\, x) = ?$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(c_2|\, x) = ?$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes

What is the probability of "Yes" and what is the probability of "No" **after observing the new data instance?**
We use Bayes' Rule for this.

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_1 \mid x) = \frac{p(x \mid c_1)p(c_1)}{p(x)}$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(c_2 \mid x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1} = \mathbf{7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2} = \mathbf{6}$

Naïve Bayes

Since we are going to choose the c_i maximizes $p(c_i \mid x)$, we can ignore $p(x)$ and only need to further compute $p(x|c_i)$ **for each class**

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1 \mid x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2 \mid x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1} = \mathbf{7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2} = \mathbf{6}$

Naïve Bayes

Since we are going to choose the c_i maximizes $p(c_i \mid x)$, we can ignore $p(x)$ and only need to further compute $p(x|c_i)$ **for each class**

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1 \mid x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2 \mid x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x \mid c_1) = p(X_1 = Sunny, X_2 = Yes, X_3 = No|c_1)$$
$$= p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

Naïve Bayes assumes that attributes are independent

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes

We make the naïve assumption that $p(X_1 = Sunny, X_2 = Yes, X_3 = No|c_1)$ is equivalent to:
$p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i} p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x \mid c_1) = p(X_1 = Sunny, X_2 = Yes, X_3 = No|c_1)$$
$$= p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x \mid c_1) = p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

$$p(X_1 = Sunny \mid c_1) = \frac{2}{7} = 0.29$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\,x) = \frac{p(x\mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\,x) = \frac{p(x\mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

$$p(x\mid c_1) = p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

$$p(X_1 = Sunny\mid c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes|\,c_1) = \frac{3}{7} = 0.43$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1 |\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2 |\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

$$p(x \mid c_1) = p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1)$$

$$p(X_1 = Sunny \mid c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes|\, c_1) = \frac{3}{7} = 0.43$$

$$p(X_3 = No \mid c_1) = \frac{2}{7} = 0.29$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x\mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x\mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$$\boldsymbol{n_1 = 7}$$

$$p(x\mid c_1) = p(X_1 = Sunny|c_1)p(X_2 = Yes|c_1)p(X_3 = No|c_1) = \left(\frac{2}{7}\right)\left(\frac{3}{7}\right)\left(\frac{2}{7}\right) = 0.035$$

$$p(X_1 = Sunny\mid c_1) = \frac{2}{7} = 0.29$$

$$p(X_2 = Yes|\, c_1) = \frac{3}{7} = 0.43$$

$$p(X_3 = No\mid c_1) = \frac{2}{7} = 0.29$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$
$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x\mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x\mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 1 (c_1): "Yes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Overcast	No	Yes	Yes

$\boldsymbol{n_1 = 7}$

$$\boldsymbol{p(x\mid c_1)p(c_1) = 0.035(0.54) = 0.0189}$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7}$$

$$p(X_2 = Yes|c_1) = \frac{3}{7}$$

$$p(X_3 = No|c_1) = \frac{2}{7}$$

$$p(x \mid c_1)p(c_1\ ) = 0.0189$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

$$p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} \qquad p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_1) = \frac{3}{7}$$

$$p(X_3 = No|c_1) = \frac{2}{7}$$

$$p(x \mid c_1)p(c_1\,) = 0.0189$$

$$p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$$\boldsymbol{n_2 = 6}$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)}$$

$$p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7}$$

$$p(X_2 = Yes|c_1) = \frac{3}{7}$$

$$p(X_3 = No|c_1) = \frac{2}{7}$$

$$p(x \mid c_1)p(c_1\,) = 0.0189$$

$$p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)}$$

$$p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7}$$

$$p(X_2 = Yes|c_1) = \frac{3}{7}$$

$$p(X_3 = No|c_1) = \frac{2}{7}$$

$$p(x \mid c_1)p(c_1) = 0.0189$$

$$p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

$$p(x \mid c_2) = \left(\frac{1}{6}\right)\left(\frac{4}{6}\right)\left(\frac{5}{6}\right) = 0.093$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} \qquad p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_1) = \frac{3}{7} \qquad p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_3 = No|c_1) = \frac{2}{7} \qquad p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

$$p(x \mid c_1)p(c_1\,) = 0.0189$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

$$\boldsymbol{p(x \mid c_2)p(c_2) = 0.093(0.46) = 0.0428}$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x\mid c_1)p(c_1)}{p(x)}$$

$$p(c_2|\, x) = \frac{p(x\mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54$$

$$p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7}$$

$$p(X_2 = Yes|c_1) = \frac{3}{7}$$

$$p(X_3 = No|c_1) = \frac{2}{7}$$

$$p(x\mid c_1)p(c_1\ ) = 0.0189$$

$$p(c_1\mid x) = 0.0189$$

$$p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

$$p(x\mid c_2)p(c_2) = 0.0428$$

$$p(c_2\mid x) = 0.0428$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$$\boldsymbol{n_2 = 6}$$

Naïve Bayes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i} \frac{p(x|c_i)p(c_i)}{p(x)}$$

$$= argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(X_1 = Sunny|c_1) = \frac{2}{7} \qquad p(X_1 = Sunny|c_2) = \frac{1}{6} = 0.17$$

$$p(X_2 = Yes|c_1) = \frac{3}{7} \qquad p(X_2 = Yes|c_2) = \frac{4}{6} = 0.67$$

$$p(X_3 = No|c_1) = \frac{2}{7} \qquad p(X_3 = No|c_2) = \frac{5}{6} = 0.83$$

$$p(x \mid c_1)p(c_1) = 0.0189 \qquad p(x \mid c_2)p(c_2) = 0.0428$$

$$p(c_1 \mid x) = 0.0189$$

$$\boxed{p(c_2 \mid x) = 0.0428}$$

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Class 2 (c_2): "No"

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Snow	Yes	No	No
Overcast	Yes	No	No

$\boldsymbol{n_2 = 6}$

Naïve Bayes for numerical attributes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

New data instance

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
2.3	7	75	?

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

Naïve Bayes for numerical attributes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i}\, p(x|c_i)p(c_i)$$

New data instance

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
2.3	7	75	?

$$p(c_1|\,x) = \frac{p(x\mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\,x) = \frac{p(x\mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x\mid c_1) = ? \qquad p(x\mid c_2) = ?$$

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

Naïve Bayes for numerical attributes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x\mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x\mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$$p(x\mid c_1) = ? \qquad p(x\mid c_2) = ?$$

Assume each class has a normal distribution with the assumption that attributes are independent

$$p(x\mid c_i) = \prod_{j=1}^{d} p(x_j|c_i) = \prod_{j=1}^{d} f(x_j|\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2)$$

Where:

$$f(x_j|\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) = \frac{1}{\sqrt{2\pi\sigma_{ij}}}\exp\left(-\frac{(x_j - \hat{\mu}_{ij})^2}{2\cdot\hat{\sigma}_{ij}^2}\right)$$

New data instance

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
2.3	7	75	?

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

Naïve Bayes for numerical attributes

$$argmax_{c_i} p(c_i|x) = argmax_{c_i}\, p(x|c_i)p(c_i)$$

$$p(c_1|\, x) = \frac{p(x \mid c_1)p(c_1)}{p(x)} \qquad p(c_2|\, x) = \frac{p(x \mid c_2)p(c_2)}{p(x)}$$

$$p(c_1) = \frac{n_1}{n} = \frac{7}{13} = 0.54 \qquad p(c_2) = \frac{n_2}{n} = \frac{6}{13} = 0.46$$

$\boldsymbol{p(x \mid c_1)} =?$ $\qquad p(x \mid c_2) =?$

$$p(x \mid c_i) = \prod_{j=1}^{d} p(x_j|c_i) = \prod_{j=1}^{d} f(x_j|\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2)$$

$$f(x_j|\hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp\left(-\frac{(x_j - \hat{\mu}_{ij})^2}{2 \cdot \hat{\sigma}_{ij}^2}\right)$$

New data instance

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
2.3	7	75	?

Inches of rain in past 2 hours	Hours of sleep in previous night	Percentage of HW completed?	Go Hiking?
0	9	80	Yes
0.5	5	90	No
1	7	95	Yes
5	7	100	Yes
0.3	8	100	Yes
0.4	4	100	No
0.1	9	27	No
0	9	50	No
0	8	100	Yes
3	10	98	Yes
6	8	95	No
2.1	8	70	No
1.02	8.5	98	Yes

Algorithm 18.2: Naive Bayes Classifier

NaiveBayes ($\mathbf{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n}$):

for $i = 1, \cdots, k$ **do**

- $\mathbf{D}_i \leftarrow \{\mathbf{x}_j \mid y_j = c_i, j = 1, \cdots, n\}$ `// class-specific subsets`
- $n_i \leftarrow |\mathbf{D}_i|$ `// cardinality`
- $\hat{P}(c_i) \leftarrow n_i/n$ `// prior probability`
- $\hat{\boldsymbol{\mu}}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ `// mean`
- $\mathbf{Z}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}_i^T$ `// centered data for class` c_i
- **for** $j = 1, .., d$ **do** `// class-specific variance for` X_j
  - $\hat{\sigma}_{ij}^2 \leftarrow \frac{1}{n_i} Z_{ij}^T Z_{ij}$ `// variance`
- $\hat{\boldsymbol{\sigma}}_i = \left(\hat{\sigma}_{i1}^2, \ldots, \hat{\sigma}_{id}^2\right)^T$ `// class-specific attribute variances`

return $\hat{P}(c_i), \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i$ *for all* $i = 1, \cdots, k$

Testing ($\mathbf{x}$ **and** $\hat{P}(c_i)$, $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\sigma}}_i$, **for all** $i \in [1, k]$):

$$\hat{y} \leftarrow \arg\max_i \left\{ \hat{P}(c_i) \prod_{j=1}^{d} f(x_j | \hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) \right\}$$

return $\hat{y}$

Naïve Bayes algorithm (numerical attributes)

Evaluating classification algorithms

CSCI 347

Adiesha Liyana Ralalage

Evaluation of Classification

Input data matrix

Class labels are always categorical

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

New data instance

Weather	Weekend?	Finished HW	Go Hiking?
Sunny	Yes	No	?

Goal is to predict the class of the new data

TRAINING SET AND TEST SET

Input data matrix

Weather	Weekend?	Finished HW	Go Hiking?
~~Snow~~	~~Yes~~	~~No~~	~~Yes~~
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes

Evaluate the class predictions of test data

TRAINING SET AND TEST SET

Input data matrix

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes

Evaluate the model using a subset of test data that you picked, which you did not use to create the model.

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

TRAINING SET AND TEST SET

Often split 80/20

Input data matrix

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	Yes	Yes	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	Yes

Evaluate the class predictions of test data based on an algorithm that built a model using only training data.

Training data

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Pipeline of evaluation

Training data

Weather	Weekend?	Finished HW	Go Hiking?
Overcast	No	No	No
Sunny	Yes	No	Yes
Overcast	No	Yes	Yes
Snow	No	Yes	No
Overcast	Yes	No	No
Sunny	Yes	No	No
Sunny	No	Yes	Yes
Snow	Yes	No	No
Overcast	Yes	No	No
Overcast	No	Yes	Yes

Give a classification algorithm training data to use to output a model

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	Yes

Feed the model unlabeled test data to make predictions

"predict the test data classes"

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	No

Compare the predicted classes to ground truth classes

Evaluation metrics

How to compare the predicted classes to ground truth classes?

Predicted classes

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	No

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	Yes

Evaluation metrics

How to compare the predicted classes to ground truth classes?

Predicted classes

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	No

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	Yes

Accuracy:

$$\frac{1}{n_T}\sum_{i=1}^{n_T} I(y_i = \hat{y}_i)$$

Where: $I(y_i = \hat{y}_i)$ is 1 if y_i and $\hat{y}_i$ have the same value, and is 0 otherwise

Evaluation metrics

How to compare the predicted classes to ground truth classes?

Predicted classes

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	No

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	Yes

Accuracy:

$$\frac{1}{n_T}\sum_{i=1}^{n_T} I(y_i = \hat{y}_i) = \frac{1}{3}(1+1+0) = \frac{2}{3}$$

In class activity

- Assuming that we have trained a classification model to predict fraudulent transactions, we used the model to predict the labels of a randomly selected test set of 10,000 transactions, which were not included in the training data. Out of these transactions, 100 were fraudulent, but the model predicted all 10,000 record as "non-fraudulent". What is the accuracy of the model based on this prediction?

Other Evaluation Metrics

How to compare the predicted classes to ground truth classes?

Predicted classes

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	No

Test data instance

Weather	Weekend?	Finished HW	Go Hiking?
Snow	Yes	No	Yes
Overcast	Yes	Yes	Yes
Snow	No	Yes	Yes

Contingency-based measures

- **Precision, recall, F-measure**

For Binary classification

- **TP, FN, FP, FN**
- **Sensitivity, specificity**

Bayes Theorem

- Estimating prior probability $P(c_i)$
  - Let $D_i = \{x_j \in D \mid x_j \; has \; class \; y_j = c_i\}$
  - $|D| = n$, and $|D_i| = n_i$
  - $\hat{P}(c_i) = \frac{|D_i|}{|D|} = \frac{n_i}{n}$
- There are two ways you can use to estimate the likelihood $P(x|c_i)$
  - Parametric approach
    - We make an assumption about the distribution of a particular class.
    - Ex: Each class is normally distributed.
  - Non-parametric approach

Bayes Theorem

- Non-parametric approach
  - We make no assumptions about the probability distribution of a class.
  - We still want to compute $P(c_i|x) = \frac{P(x|c_i) \cdot P(c_i)}{P(x)}$
  - We can still estimate the $P(c_i)$, just like we did earlier.
  - How do we calculate $P(c_i|x)$ without making any assumptions about the distribution of the class?
  - One approach is to use k-nearest neighbor density estimation.
  - This algorithm is quite simple.

Bayes Theorem

- Non-parametric approach
  - For a given point we look at a radius r hypersphere such that we include its k-nearest neighbor.
  - We can use any distance measurement for this radius, e.g., $L_1, L_2, \ldots, L_\infty$.
  - We can count the number of points in each class in this hypersphere.
  - We can use this information to calculate $P(c_i|x)$
    - $P(c_i|x) = \frac{\#\ of\ neighbors\ with\ class\ c_i}{k}$
  - This is called a lazy classifier
  - You can try a small value for k and check the accuracy of the classification.
  - There is no way to find the best possible k value theoretically.
  - Surprisingly, this works very well on real-world datasets.

Bayes Theorem

- Non-parametric approach
  - How can we tackle ties?
    - Simplest solution is to make k an odd value (if there are only two classes)
      - For multi-class problem still, we could have this problem.
    - Use weighted voting
      - Instead of simply counting votes, assign higher weights to closer neighbors.
        - $w_i = \frac{1}{d_i} : d_i \; is \; the \; distance \; to \; the \; i^{th} \; neighbor$
        - Class with the highest total weight wins.
    - Random selection
      - If there is a tie, pick one of the tied class randomly.
    - Prioritize classes based on frequency
      - Requires prior knowledge of class distributions

Evaluation methods for classification

- In the last lecture we looked at $\text{Accuracy} := \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$
  - Basically, count the number of mismatches and sum them up, and take the arithmetic mean.
- $ErrorRate = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) = 1 - Accuracy$
- We can also look at the Precision, Recall and F-score of the output as well which are contingency table-based methods.
- $prec_i = \frac{n_{ii}}{m_i}$ $recall_i = \frac{n_{ii}}{n_i}$ $F_i = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i}$ $F = \frac{1}{k}\sum_{i=1}^{k} F_i$

MONTANA STATE UNIVERSITY

Binary classification

When you only have two classes

- We call the class c_1 as the positive class and class c_2 as the negative class.
  - What is the size of the confusion matrix (a.k.a contingency table)?
- $True\ Positives\ (TP)$: the number of points that the classifier correctly predicts as positive
  - $TP = n_{11} = |\{x_i \mid \hat{y}_i = y_i = c_1\}|$
- $False\ Positives\ (FP)$: the number of points classifier incorrectly predicts as positives where they are negatives.
  - $FP = n_{12} = |\{x_i | \hat{y}_i = c_1\ and\ y_i = c_2\}|$
- $False\ Negatives$: the number of points incorrectly predicted as negatives where they are positives:
  - $FN = n_{21} = |\{x_i\ | \hat{y}_i = c_2\ and\ y_i = c_1\}|$
- $True\ Negatives$: the number of points correctly predicted as negatives
  - $TN = n_{22} = |\{x_i\ | \hat{y}_i = y_i = c_2\}|$

	True Class	
Predicted Class	Positives (c_1)	Negatives (c_2)
Positives (c_1)	True Positives (TP)	False Positives (FP)
Negatives (c_2)	False Negatives (FN)	True Negatives (TN)

Contingency table/Confusion matrix for Two classes

$$\boldsymbol{Error\ Rate = \frac{FP + FN}{n}}$$

$$\boldsymbol{Accuracy\ = \frac{TP + TN}{n}}$$

These are global measures of classifier performance.

MONTANA STATE UNIVERSITY

	True Class	
Predicted Class	Positives (c_1)	Negatives (c_2)
Positives (c_1)	True Positives (TP)	False Positives (FP)
Negatives (c_2)	False Negatives (FN)	True Negatives (TN)

Contingency table/Confusion matrix for Two classes

$$\boldsymbol{prec_P = \frac{TP}{TP + FP} = \frac{TP}{m_1}}$$

$$\boldsymbol{prec_N = \frac{TN}{FN + TN} = \frac{TN}{m_2}}$$

These are class specific Precision values

	True Class	
Predicted Class	Positives (c_1)	Negatives (c_2)
Positives (c_1)	True Positives (TP)	False Positives (FP)
Negatives (c_2)	False Negatives (FN)	True Negatives (TN)

Contingency table/Confusion matrix for Two classes

Sensitivity is called True Positive Rate

$$TPR = recall_P = \frac{TP}{TP + FN} = \frac{TP}{n_1}$$

Specificity is called True Negative Rate

$$TNR = recall_N = \frac{TN}{FP + TN} = \frac{TP}{n_2}$$

False Negative Rate

$$FNR = \frac{FN}{TP + FN} = \frac{FN}{n_2} = 1 - sensitivity$$

False Positive Rate

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{n_2} = 1 - sepecificity$$

MONTANA STATE UNIVERSITY

Classifier Evaluation

- We can also evaluate a classifier M using some performance measure θ.
- We use something called $K - Fold\ cross\ validation$

K-fold cross validation

Idea

- We divide the dataset D into K equal sized parts, called folds, namely $D_1, D_2, \dots, D_k$.
- Each fold D_i is treated as a testing set, with remaining folds comprising the training set $D \setminus D_i = \bigcup_{j \neq i} D_j$
- After training the dataset, we asses the performance of the D_i testing set and retrieve θ_i
- Then we calculate the $\hat{\mu}_\theta = E[\theta] = \frac{1}{K}\sum_{i=1}^{K} \theta_i$ and variance $\hat{\sigma}_\theta^2 = \frac{1}{K}\sum_{i=1}^{K}(\theta_i - \hat{\mu}_\theta)^2$

MONTANA STATE UNIVERSITY