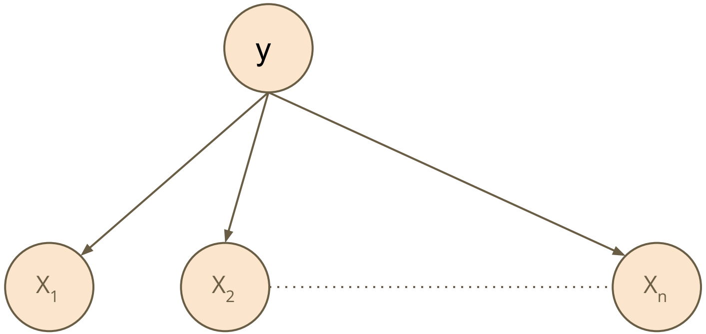
Probabilistic Methods
Fall 2024

Contents

- Probabilistic Methods - Naive Bayes Classifier
    - Discrete and Continuous Cases
- How to handle non-numeric data? Mix of types?
- How to handle missing data?
- Bias vs Variance

Naive Bayes Classifier

- It works by estimating probabilities
- The prediction variable is y, and the features $X_1, X_2, \ldots, X_3$
- NBC learns a Naive bayesian model
- Features in the dataset are IID, independent and identically distributed.



https://www.flickr.com/photos/mattbuck007/3676624894

The decision rule

Consider a binary classifier, with classes A, B (values of label, **y**)

If $Pr(y=A|X_1=v_1,X_2=v_2,\ldots_1,X_n=v_n) > Pr(y=B|X_1=v_1,X_2=v_2,\ldots,X_n=v_n)$

Predict $y = A$

Else

Predict $y = B$

An example

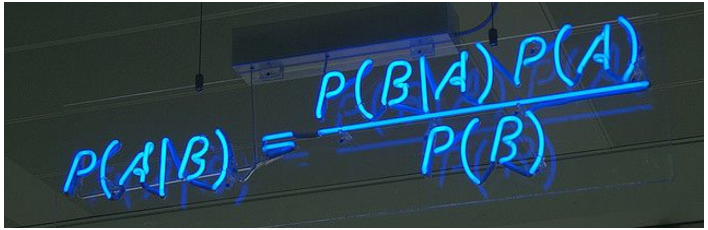
Pr(PlayGolf=yes | **Outlook=Overcas** **Temp=Mild, Humidity=Normal, Windy = True**) = ?

Pr(PlayGolf=no | **Outlook=Overcast** **Temp=Mild, Humidity=Normal, Windy = True**) = ?

Predictors Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example



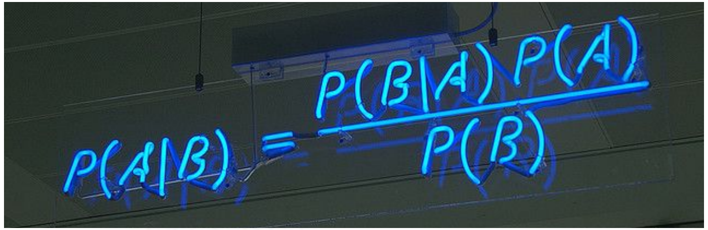
Pr(PlayGolf=yes | **Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**) = ?

= Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True** | PlayGolf=yes) * Pr(PlayGolf=yes) / Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**)

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example



Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=yes)

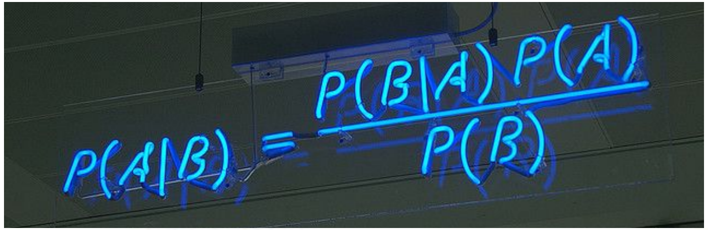
Pr(PlayGolf=yes)

Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**)

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example



Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=yes) =

Pr(**Outlook=Overcast,**|PlayGolf=yes) *

Pr(**Temp=Mild**|PlayGolf=yes) *

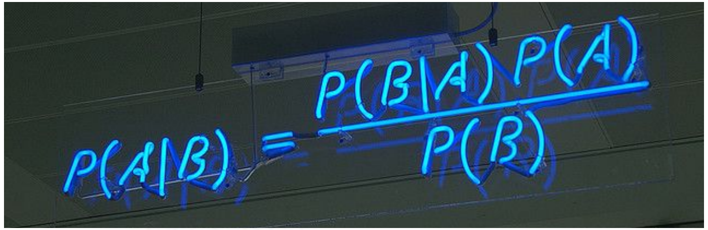
Pr(**Humidity=Normal**PlayGolf=yes) *

Pr(**Windy = True**|PlayGolf=yes)

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example



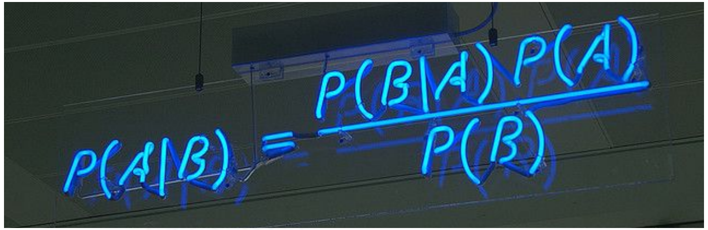
Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=yes) =
(4/9) * (4/9) * (6/9)*(3/9)

Pr(**Outlook=Overcast,**|PlayGolf=yes)
= 4/9
Pr(**Temp=Mild**|PlayGolf=yes)
= 4/9
Pr(**Humidity=Normal**PlayGolf=yes)
= 6/9
Pr(**Windy = True**|PlayGolf=yes)
= 3/9

Predictors / Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example



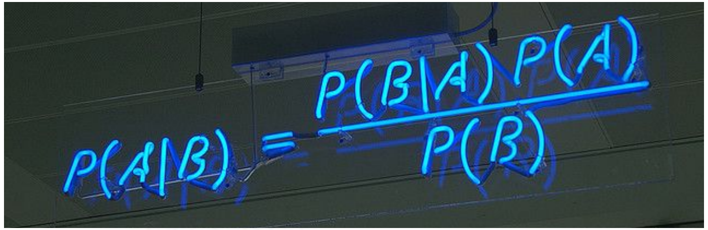
Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=yes) = (4/9) * (4/9) * (6/9)*(3/9)

Pr(PlayGold=yes) = 9/14

Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**) = ?

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example



Pr(PlayGolf=no|**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**) = ?

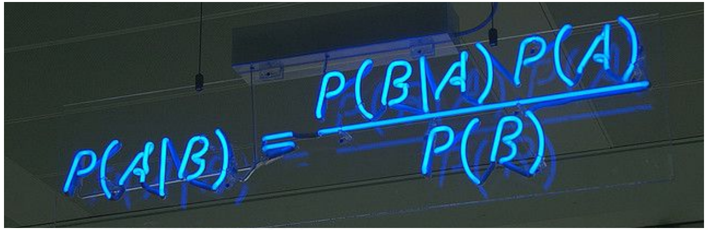
= Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=no) * Pr(PlayGold=no) / Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**)

Predictors				Target
Outlook	**Temp**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example

Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=no) =
(0/5) * (2/5) * (1/5)*(3/5)

Pr(**Outlook=Overcast,**|PlayGolf=no)
= 0/5
Pr(**Temp=Mild**|PlayGolf=no)
= 2/5
Pr(**Humidity=Normal**|PlayGolf=no)
= 1/5
Pr(**Windy = True**|PlayGolf=no)
= 3/5



Predictors | Target

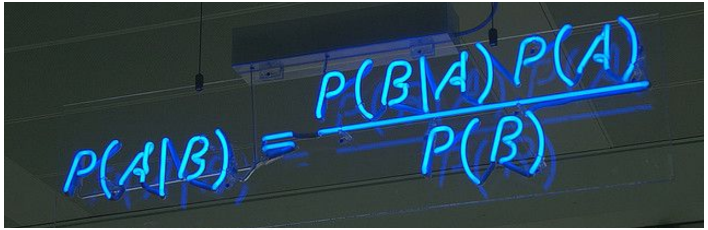
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

An example

Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**|PlayGolf=no) = (0/5) * (2/5) * (1/5)*(3/5)

Pr(PlayGold=yes) = 5/14

Pr(**Outlook=Overcast, Temp=Mild, Humidity=Normal, Windy = True**) = ?



Predictors				Target
Outlook	**Temp**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

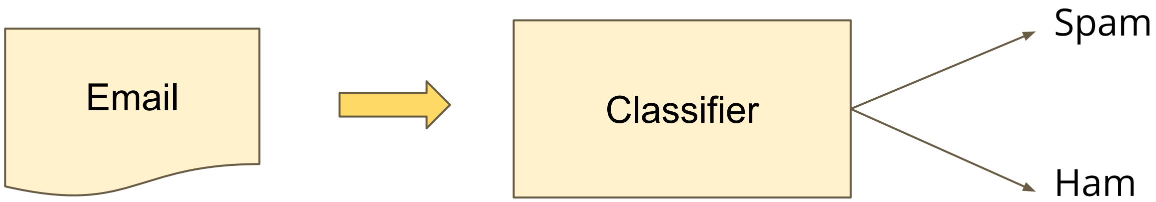
Pros and Cons

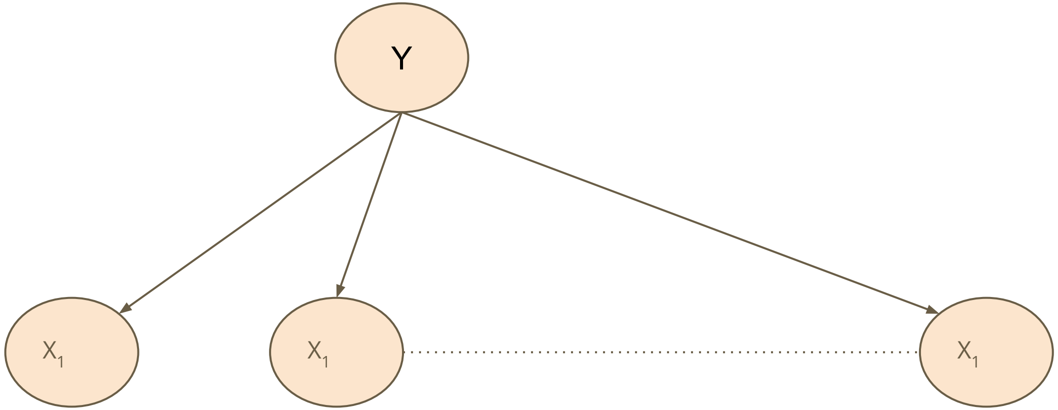
- Easy to implement
- Training faster, no gradient descent needed
- Does not overfit
- Less memory and CPU requirement
- Easy to retrain with new data
- Handles both numeric and categorical data
- Handles missing data automatically

- Ensembling, bagging, boosting does not work, no variance to reduce
- Zero frequency problem for categorical data
- IID assumption - often not practical
- Numerical underflow
- Skewed data causes problems in prior calculation

Another Example - Text Classification

- Spam Filter
- Bag of words





Supervised Learning Problem

- Bag of words model
- Features are frequency of the words
- Maintains a dictionary of words == vocabulary

SPAM

OFFER IS SECRET
SECRET SPORTS LINK
CLICK SECRET LINK

HAM

WENT PLAY SPORTS
PLAY SPORTS TODAY
SECRET SPORTS EVENT
SPORTS IS TODAY
SPORTS COSTS MONEY

Email Filter

- Vocabulary = 12
- P (SPAM) = ?
- P(“SECRET” | SPAM) = ?
- P(“SECRET” | HAM) = ?
- P(“SPORTS” | SPAM) = ?
- P(“SPORTS” | HAM) = ?

SPAM

OFFER IS SECRET
SECRET **SPORTS LINK**
CLICK SECRET LINK

HAM

WENT PLAY SPORTS
PLAY SPORTS **TODAY**
SECRET SPORTS **EVENT**
SPORTS IS TODAY
SPORTS **COSTS MONEY**

Email Filter

SPAM

OFFER IS SECRET
SECRET **SPORTS LINK**
CLICK SECRET LINK

HAM

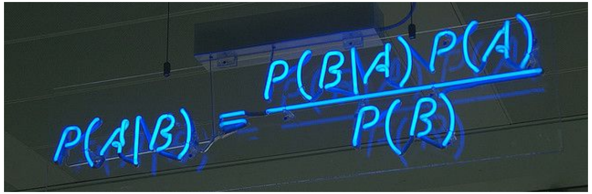
WENT PLAY SPORTS
PLAY SPORTS **TODAY**
SECRET SPORTS **EVENT**
SPORTS IS TODAY
SPORTS **COSTS MONEY**

Message = “SPORTS”

If **Pr (SPAM|Message = “SPORTS“)** > **Pr (HAM|Message = “SPORTS“)**
    Message = **SPAM**
Else
    Message = **HAM**

Pr (SPAM | SPORTS) = Pr (SPORTS|SPAM) * Pr (SPAM) / Pr (SPORTS)

Pr (HAM | SPORTS) = Pr (SPORTS|HAM) * Pr (HAM) / Pr (SPORTS)



Email Filter

SPAM

OFFER IS SECRET
SECRET **SPORTS LINK**
CLICK SECRET LINK

HAM

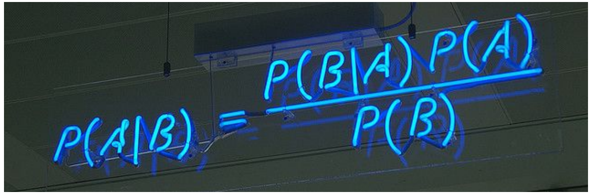
WENT PLAY SPORTS
PLAY SPORTS **TODAY**
SECRET SPORTS **EVENT**
SPORTS IS TODAY
SPORTS **COSTS MONEY**

Message = "SECRET IS SECRET"

Pr (SPAM | MESSAGE) = Pr (MESSAGE|SPAM) * Pr (SPAM) / Pr (MESSAGE)

Pr (HAM | MESSAGE) = Pr (MESSAGE|HAM) * Pr (HAM) / Pr (MESSAGE)

Pr (MESSAGE) = Pr (MESSAGE|SPAM) * Pr (SPAM) + Pr (MESSAGE|HAM) * Pr (HAM)



Email Filter

Message = "TODAY IS SECRET"

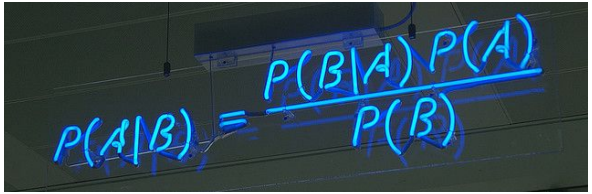
SPAM

OFFER IS SECRET
SECRET **SPORTS LINK**
CLICK SECRET LINK

HAM

WENT PLAY SPORTS
PLAY SPORTS **TODAY**
SECRET SPORTS **EVENT**
SPORTS IS TODAY
SPORTS **COSTS MONEY**

Pr (SPAM | MESSAGE) = Pr (MESSAGE|SPAM) * Pr (SPAM) / Pr (MESSAGE)
Pr (HAM | MESSAGE) = Pr (MESSAGE|HAM) * Pr (HAM) / Pr (MESSAGE)
Pr (MESSAGE) = Pr (MESSAGE|SPAM) * Pr (SPAM) + Pr (MESSAGE|HAM) * Pr (HAM)



Laplacian Smoothing

- Pr (X) = Count (x) / Total Count
- In Laplacian Smoothing,
  - Pr (X) = (Count (X) + k) / (Total Count + K * |x|)
- P (SPAM) = ?
- P (“TODAY” | SPAM) = ?
- P (“IS”|SPAM) = ?
- P (“SECRET”|SPAM) = ?
- P (“TODAY” | HAM) = ?
- P (“IS”|HAM) = ?
- P (“SECRET”|HAM) = ?

SPAM

OFFER IS SECRET
SECRET **SPORTS LINK**
CLICK SECRET LINK

HAM

WENT PLAY SPORTS
PLAY SPORTS **TODAY**
SECRET SPORTS **EVENT**
SPORTS IS TODAY
SPORTS **COSTS MONEY**

Message = “TODAY IS SECRET”

Advanced Email Filters

- Known spamming IP
- Have you emailed the person before?
- Other people received the same message?
- Email header consistent?
- ALL CAPS?
- URLs are pointing correctly?
- Are addressed by your correct name?

Digit Recognition

- Features = Pixels
- Class Labels = 0,1,...,9
- Lets use sklearn

Gaussian Naive Bayes

- How to handle numeric data?
- We assume that these values are sampled from a gaussian distribution

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

$$P(Income = 120 \mid Evade = No) = \frac{1}{\sqrt{2 \times \pi \times 2975}} \exp\left(-\frac{(120 - 110)^2}{2 \times 2975}\right)$$

$$P(Income = 120 \mid Evade = Yes) = \frac{1}{\sqrt{2 \times \pi \times 25}} \exp\left(-\frac{(120 - 90)^2}{2 \times 25}\right)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Bias vs Variance

Voting Republican	Voting Democratic	Non-Respondent	Total
13	16	21	50

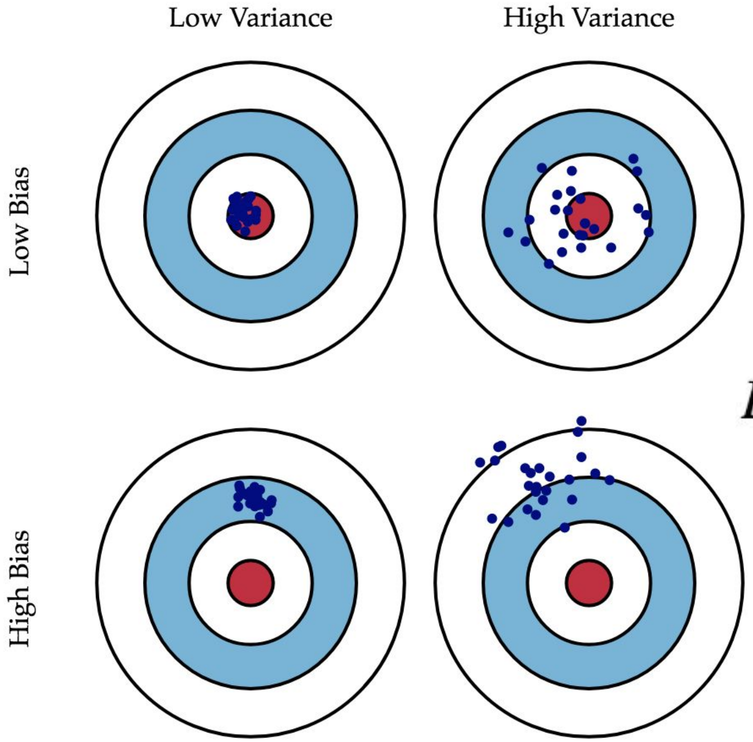
- Bias are the simplifying assumptions made by a model to make the target function easier to learn.
- Suppose you want to find the number of votes that Joe Biden will get in different states
  - The real value is y
  - Your prediction is $\hat{f}(x)$
  - The bias is the difference

$$bias = E[\hat{f}(x)] - f(x)$$

- Variance is the expectation of the squared deviation of a random variable from its mean
  - Estimate of the target function will change if different training data was used.

$$var(x) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

Bias vs Variance



$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias vs Variance

Low Bias

Depends on Training Data, Not much assumption on the model, KNN, SVM, Decision Tree

High Bias

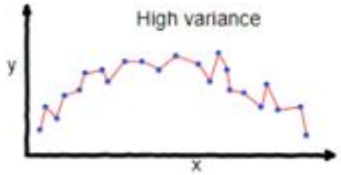
Assumptions on the model, underfitting, not adequate data, linear and logistic regression

Low Variance

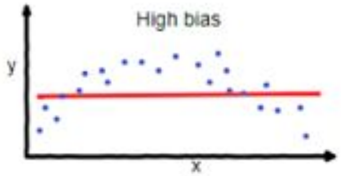
Changing the dataset makes small changes on the model

High Variance

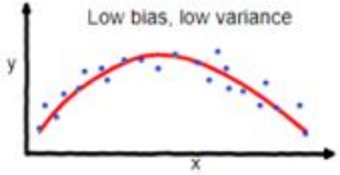
Depends on training data, captures noise, overfits



overfitting

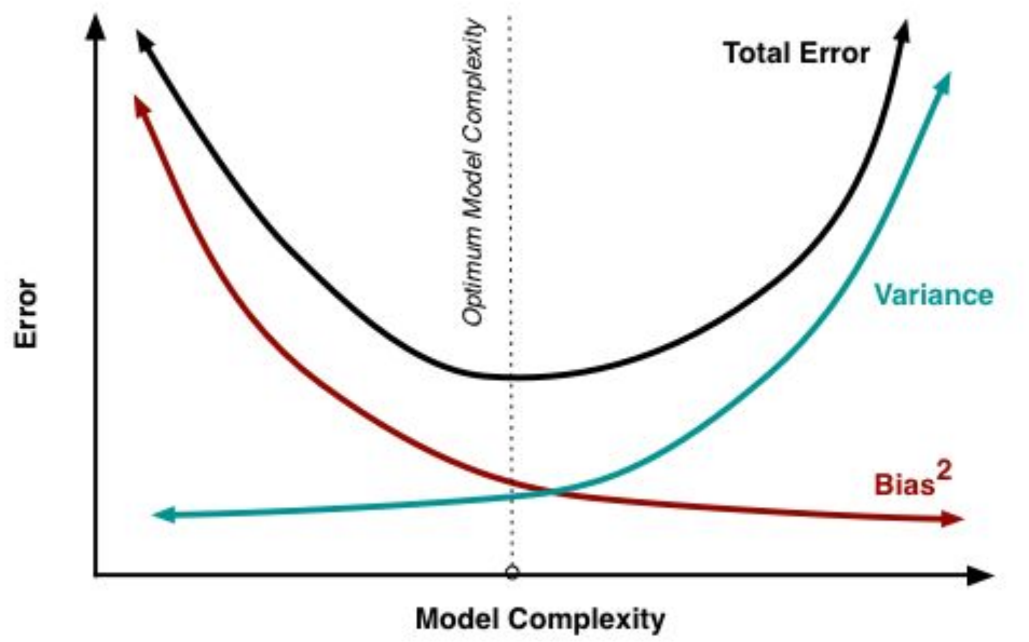


underfitting



Good balance

Bias vs Variance



$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

http://scott.fortmann-roe.com/docs/BiasVariance.html