Տվյալների բաժանում: Գերուսուցում, թերուսուցում

Հայկ Կարապետյան

1 Տվյալների բաժանում

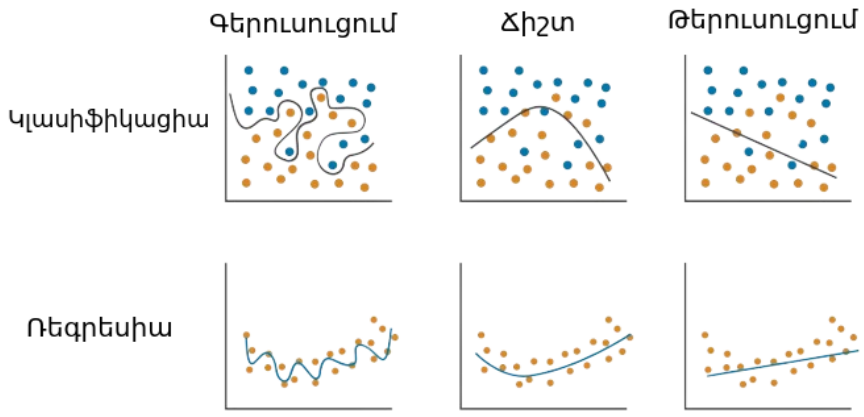
Որպեսզի ստուգենք մեր մոդելը[1] կարողանում է ճիշտ գուշակել պիտակը ոչ միայն մեր ունեցած տվյալների համար այլև մեր չունեցած տվյալների համար, մեր տվյալները բաժանում ենք 3 մասի՝

1. train data (ուսուցման տվյալներ)
2. test data (փորձարկման տվյալներ)
3. validation data (կարգավորման տվյալներ)

Ուսուցման տվյալները օգտագործվում են մոդելին սովորեցնելու համար: Փորձարկման տվյալների միջոցով կարող ենք հասկանալ, մոդելը միայն ուսուցման տվյալների համար է լավ գուշակություն կատարում, թե ոչ միայն: Կարգավորման տվյալների միջոցով կարող ենք փոփոխել մոդելի որոշակի արժեքներ: Հիմնականում տվյալների բաժանումը տեղի է ունենում պատահական կերպով: Տվյալներից պատահական կերպով ընտրվում է 70% ուսուցման տվյալներ, 15% փորձարկման տվյալներ, 15% կարգավորման տվյալներ:

2 Գերուսուցում, թերուսուցում

Գերուսուցումը (overfitting) և թերուսուցումը (underfitting) մոդելի վիճակներ են: Կասենք որ մոդելը գերուսուցված է, երբ ուսուցման տվյալների վրա մոդելի արդյունքը լավն է (ճիշտ է գուշակում պիտակները), իսկ փորձարկման տվյալների վրա լավը չէ: Այսինքն մեր ուսուցման տվյալները անգիր ենք արել: Կասենք որ մոդելը թերուսուցված է, երբ ո՛չ ուսուցման տվյալների, ո՛չ էլ փորձարկման տվյալների վրա մոդելի արդյունքը լավը չէ:



Նկար 1: Մոդելի վիճակներ՝ կախված ցուցաբերած արդյունքից

1. Մոդելը մեքենայական ուսուցման այն ալգորիթմն է, որը ուսուցանելու ենք մեր տվյալներով