

Python för AI, inlämning projekt del 1.

Det egna projektet, steg 1. Mitt problem.....	1
ANGREPPSSÄTT / METOD.....	1
DATA	2

Det egna projektet, steg 1. MITT PROBLEM

Jag har ambitionen att angripa en uppgift om vädret i närområdet Stockholm med omnejd.

Jag bor nära Bromma flygplats, men jag befinner mig ofta vid Tullinge på fritidsaktivitet i en terräng som oftast är lite kallare på vintern, samt på Värmdö nära skärgården.

Min tanke är att samla data från ett antal stationer i närområdet – med utgångspunkt från Stockholm city som har en lång obruten serie, och försöka skapa en ML/AI-tolk som ger slutsatser om vädret/en väderparameter på en punkt givet data från andra punkter.

Jag bedömer att det ska finnas bra med data, i alla fall rörande parametrar som temperatur, nederbörd, vind - kanske för molnighet och fler. Vad det gäller andra parametrar är det mer osäkert men det kan vara relevant och det gör frågeställningen mer intressant om man kan bedöma saker även utifrån andra.

Jag har ännu inte landat i exakt vilken parameter jag vill fokusera på – det kan vara temperaturens avvikelse på grund av orografi eller närhet till vatten – eller så försöker jag se om man kan hitta samband t ex när moln eller eventuella skurar uppstår på tidig sommar då sjöbrisens sätter igång från havet.

ANGREPPSSÄTT / METOD

Jag tänker mig att jag ska angripa detta som ett regressionsproblem. Jag har ett ungefärligt hum om hur terräng, närhet till havet, beroendet av säsong samt vindriktning påverkar. Så jag kan träna modellen mot rimliga svar – och sedan justera fram faktorer förnumeriska relationer, bias etc.

Jag kommer att laborera enligt följande.

Steg 1: Börja med att hitta temperatursamband, där t ex Tullinge ligger på en plats som brukar ha 1-3 grader kallare än Bromma. Och Värmdö ha mer påverkan av havet med dess säsongsvariation och även tröghet.

Steg 2: Sedan även utforska om jag kan hitta samband kopplat till molnighet och nederbörd. Detta är troligen svårare, men det skulle t ex vara intressant att hitta korrelationer mellan vind, molnbildning och eventuella skurar nu inför sommarens sjöbrissäsong. Under denna (maj-juli främst) uppvisar vind, molnighet och nederbörd speciella mönster kopplat till tid och plats under dagar med sjöbris.

Jag har således både fysikalisk/meteorologisk kunskap för att laborera de olika dataseten mot varandra – och dataset med volym och kvalitet enligt nedan.

DATA (allmänt om urvalet)

Källor

Temperaturen.nu

Stort antal lokala stationer, med varierande kvalitet. Eftersom källaren är privatpersoner och därmed inte kvalitetssäkrad, så får man titta på datan för att bedöma om de dels verkar rimliga och dels har tillräcklig kvalitet.

Data finns 2,5 år bakåt, dygns-medelvärden och tim-medelvärden.

SMHI

En STOR mängd data med olika väderparametrar från lång tillbaka i tiden.

Ladda ner meteorologiska observationer

Nederbördsmängd (dygn): SMHIs stationsnät		<input type="text" value="Sök station"/> Sök
VÄLJ PARAMETER	FILTRERA STATIONER	VISA STATIONSLISTA
Lufttemperatur (h)	Lufttemperatur (dygn)	Lufttemperatur (månad)
Lufttemperatur, min och max (12h)	Lufttemperatur, min och max (dygn)	Daggpunktstemperatur (h)
Nederbördsmängd (15 min)	Nederbördsmängd (h)	Nederbördsmängd (dygn) <input checked="" type="checkbox"/>
Nederbördsmängd (månad)	Nederbördssintensitet (15 min)	Nederbördssintensitet, max av medel (15 min)
Nederbördstyp (12h)	Nederbördstyp (dygn)	Snödjup och markytans tillstånd (dygn)
Relativ luftfuktighet (h)	Vindriktning och vindhastighet (h)	Vindhastighet, max av medel (h)
Bywind, max (h)	Total molnmängd (h)	Signifikanta moln (h)
Lägsta molnbas (h)	Lägsta molnbas, min (15 min)	Solskenstid (h)
Globalstrålning (h)	Långvågsstrålning (h)	Lufttryck (h)
Sikt (h)	Rådande väder (h)	Fråga oss

Till exempel:

Nederbörd

Tillbaka till 1949-01-01. Dygnsvärdet. 23500 rader i csv.

Temperaturer

På den äldsta stationen, Stockholm-Observatoriekullen, data ända från 1859-01-01

Datapunkter

De tre områdena är de jag vill hitta korrelation/kunna förutsäga inbördes är i de röda ringarna – grunddata med längsta mätserien i den blå hexagonen.



Data, analys

- Är det komplett?

För flygplatser brukar det finnas obrutna långa serier, även för SMHIs officiella mätdata. Och det är möjligt att det även finns det för andra mätpunkter. En första analys visar på att det ser komplett ut.

- Har du null-värden?

Väderdata kan nog alltid ha luckor i datamängden på grund av tekniska problem och annat. Antingen har man korrigerat detta genom att extrapolera eller göra medelvärde, eller så får jag angripa det själv och göra en lösning.

Exempel Temperaturen.nu

	A	B	C	D
1	Datum	Klockslag	Timmedel	
2	2021-09-15	14:00:00	15,9	
3	2021-09-15	15:00:00	15,6	
4	2021-09-15	16:00:00	15,9	
5	2021-09-15	17:00:00	14,5	
6	2021-09-15	18:00:00	13,9	
7	2021-09-15	19:00:00	12,7	
8	2021-09-15	20:00:00	11,3	
9	2021-09-15	21:00:00	9,7	

En första analys visar på att NULL/luckor är korrigerade.

Exempel SMHI

Som framgår är datat kvalitetssäkrat.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Stationsnamn	Stationsnummer	Stationsnät	Måthöjd (meter över marken)									
2	Tullinge A		97100	SMHIs statio 2.0									
3													
4	Parameternamn	Beskrivning	Enhet										
5	Nederbördsmängd	summa 1 dygn, 1 går millimeter											
6													
7	Tidsperiod (fr.o.m)	Tidsperiod (t.o.m)	Höjd (meter i Latitud (deci)	Longitude (decimalgrader)									
8	1949-01-01 00:00	1969-11-02 23:59	45.0	59.1833	17.9167								
9	1969-11-01 00:00	1985-07-01 23:59	54.0	59.1833	17.9167								
10	1995-12-01 00:00	2024-04-22 12:20	44.858	59.1785	17.9093								
11													
12	Från Datum Tid (UTC)	Till Datum Tid (UTC)	Representati	Nederbördsn	Kvalitet		Tidsutsnitt:						
13	2023-12-13 06:00	2023-12-14 06:00	2023-12-13 0.1	G			Data från senaste fyra månaderna						
14	2023-12-14 06:00	2023-12-15 06:00	2023-12-14 0.1	G			Tidsperiod (fr.o.m.) = 2023-12-14 00:00:01 (UTC)						
15	2023-12-15 06:00	2023-12-16 06:00	2023-12-15 0.0	G			Tidsperiod (t.o.m.) = 2024-04-22 12:00:00 (UTC)						
16	2023-12-16 06:00	2023-12-17 06:00	2023-12-16 0.0	G			Värdet är summerat över = 24 timmar						
17	2023-12-17 06:00	2023-12-18 06:00	2023-12-17 1.5	G			Kvalitetsskoderna:						
18	2023-12-18 06:00	2023-12-19 06:00	2023-12-18 0.4	G			Grön (G) = Kontrollerade och godkända värden.						
19	2023-12-19 06:00	2023-12-20 06:00	2023-12-19 0.0	G			Gul (Y) = Misstänkt eller aggregerade värden. Grovt kontrollerade arkivdata och okontrollerade Nätinformation.						
20	2023-12-20 06:00	2023-12-21 06:00	2023-12-20 3.4	G			SMHIs Stationsnät: Data samlas in och lagras i SMHIs databaser. Data kvalitetsskoderas.						
21	2023-12-21 06:00	2023-12-22 06:00	2023-12-21 9.7	G			Övriga stationer: Data samlas in och lagras i SMHIs databaser. Dataläget är för SMHI och						
22	2023-12-22 06:00	2023-12-23 06:00	2023-12-22 5.6	G			Möjliga orsaker till saknade data:						
23	2023-12-23 06:00	2023-12-24 06:00	2023-12-23 1.0	G			#NAN?						
24	2023-12-24 06:00	2023-12-25 06:00	2023-12-24 3.1	G			- stationen har endast levererat värden med kvalitetsskod Röd (R). Dessa levereras ej.						
25	2023-12-25 06:00	2023-12-26 06:00	2023-12-25 1.9	G									
26	2023-12-26 06:00	2023-12-27 06:00	2023-12-26 0.0	G									
27	2023-12-27 06:00	2023-12-28 06:00	2023-12-27 0.0	G									

- Har du extrema värden?

Är det officiella data brukar den vara tvättad eller korrigerad. Men givetvis måste jag kontrollera detta och vid behov skapa en funktion för att korrigera.

- Vilka datatyper har datat?

Vad gäller temperatur och nederbörd till exempel lär det vara decimaltal. (Se csv-fil kopia ovan). Vad det gäller vind måste man tänka på att man dels har vindriktning, och utöver medelvinden även kan ha byvind – det kan kräva en numerisk ansats för att göra om till float och kategorier. Vad gäller molnighet kan det vara relevant hur stor andel av himmeln som har moln och kanske även vilken höjd, för detta finns det olika angreppssätt beroende om det kommer från en flygplats eller från en vanlig väderstation.

- Vilka fält i ditt data vill du använda dig av?

Fälten med data för parametrarna ovan, och tiden kopplat till dem.

- Hur kan du konvertera alla fält du vill använda till ett numeriskt format?

Jag antar att man kan skapa en ny lista med uträknade värden utifrån en befintlig kolumn.