데이터과학 프로세스(4): 탐험적 데이터 분석

> *진정한 발견은 새로운 장소를 찾는 것이 아니라, 새로운 관점을 갖는 것이다.*
> *— 마르셀 프루스트*

여러분의 컴퓨터에 이제 막 수집된 데이터가 도착했다. 방금 데이터에 대한 기본적인 점검도 마쳤다. 그럼 여러분은 당장 여러분이 좋아하는 분석 프로그램을 열고 자신이 문제정의 단계에서 세웠던 여러 가설에 대한 결론을 도출해보고 싶을 것이다. 뭔가 예측하는 것이 목적이었다면 당장 기계학습 알고리즘을 적용해보고 싶을 것이다.

하지만 경험많은 데이터 과학자라면 그렇게 하지 않을 것이다. 그들은 어떤 결론에 도달하기 전에 주어진 데이터의 모든 측면을 철저히 이해하려고 노력할 것이다. 데이터 수집 과정에서 세운 모든 가정이 맞는지, 혹시 기대하지 않았던 새로운 패턴이 발견되지 않는지, 통계적 추론 및 예측 모델을 만들때 고려 사항에는 어떤 것이 있을지 등등을 알고싶어 할 것이다.

이처럼 주어진 데이터를 다양한 각도에서 들어다보고 좀더 잘 이해하기 위해 노력하는 과정을 탐험적 데이터 분석(Exploratory Data Analysis 이하 EDA)라고 부른다. '탐험적'이라는 수식어는 문자 그대로 이를 통해 어떤 것을 발견하게 될지를 미리 예측할 수 없기 때문이다. 박스플롯(boxplot)을 비롯한 수많은 업적을 남긴 통계학의 대가인 존 터키(John Tukey)는 탐험적 데이터 분석이 탐정의 일과 비슷하다고도 말했다.

탐험적 데이터 분석이 필요한 이유는 몇가지가 있다. 우선 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 더 잘 이해하고, 데이터 준비 단계에서 놓쳤을수도 있는 잠재적인 문제를 발견할 수 있다. 또한, 데이터를 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 미처 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 추가할 수 있다. 데이터에 대한 이런 지식들은 이후에 통계적 추론이나 예측 모델을 만들때 그대로 사용된다.

어떤 특정한 결론을 결론을 도출하기 위해서가 아니라, 데이터에서 최대한 다양한 이야깃거리를 뽑아내려 한다는 측면에서 탐험적 데이터 분석은 지도없이 떠나는 여행이다[1]. 그리고 작업의 특성상 탐험적 데이터 분석의 과정은 명확한 성공 요건이나 절차를 정의하기가 힘들다. 하지만 탐험적 데이터 분석을 위해 거쳐야 할 최소한의 몇가지 단계가 있다. 여기서는 이 단계에 대해 알아보도록 하자.

탐험적 분석의 단계

이제 탐험적 분석의 과정을 단계별로 살펴보자. 탐험적 분석의 자연스러운 출발점은 주어진 데이터의 각 측면에 해당하는 개별 속성의 값을 관찰하는 것이다. 개별 속성에 대한 분석이 이루어진 후에는 속성간의 관계에 초점을 맞추어 개별 속성 관찰에서 찾아내지 못했던 패턴을 발견할 수도 있다. 그리고 이런 절차는 데이터에서 흥미있는 패턴이 발견될때까지 (혹은 더이상 찾는 것이 불가능하다고 판단될 때까지) 반복된다.

탐험적 데이터 분석의 주된 수단을 살펴보자. 우선 원본 데이터를 관찰하는 방법, 다양한 요약 통계값 (statistics)을 사용하는 방법, 마지막으로 적절한 시각화를 사용하는 방법이 있다. 원본 데이터 관찰은 데이터 각 항목과 속성값을 관찰하기 때문에 꼼꼼한 반면 큰 그림을 놓치기 쉽다. 반면에 요약 통계값이나 시각화를 사용하면 숲은 보지만 나무는 보지 못하는 우를 범할 수 있다. [2]

따라서 중요한 것은 이 세가지 방법이 보완적으로, 그리고 순환적으로 사용되어야 한다는 것이다. 여기서 순환적이라는 말은 원본 데이터를 보다가 의심가는 부분이 있으면 적절한 시각화나 통계값을 통해 검증하고, 반대로 시각화나 통계값을 통해 발견한 패턴은 해당하는 원본 데이터 값을 찾아 추가적인 검증을 해야 한다는 뜻이다. 미지의 땅을 탐사할 때, 항공 정찰과 함께 실제로 그 땅에 들어가 탐사하는 과정이 모두 이루어져야 하는 것과 같은 원리다.

개별 속성 분석하기

우선 탐험적 데이터 분석의 첫번째 단계로 개별 속성을 살펴보자. 이를 통해 데이터를 구성하는 각 속성의 값이 우리가 예측한 범위와 분포를 갖는지, 만약 그렇지 않다면 왜 그런지를 알아볼 수 있을 것이다. 또한 데이터에는 다양한 이유로 정상 범주를 벗어난 값이 존재할 수 있는데, 이런 이상값(outlier)을 찾아내는 것도 탐험적 데이터 분석에서 이루어져야 한다.

개별 속성의 값을 살펴보는 방법에는 앞에서 밝힌대로 개별 데이터 관찰, 그리고 통계값 및 시각화를 활용하는 방법이 있다. 우선, 개별 데이터의 값을 눈으로 보면서 전체적인 추세와 어떤 특이사항이 있는지 관찰할 수 있다. 데이터가 작은 경우 전체를 다 살펴볼 수 있겠지만, 데이터의 양이 많은 경우 이는 시간이 많이 소요되는 일이다.

하지만 시간이 없다고 큰 데이터의 앞부분만 보는 것은 피해야 한다. 데이터 앞부분에서 나타나는 패턴과 뒷부분에서 나타나는 패턴이 상이할 수 있기 때문이다. 이런 경우 데이터에서 무작위로 표본을 추출한 후에 관찰하는 것이 올바른 방법이다. 무작위로 추출한 표본 안에는 데이터 전체가 고루 반영되어 있기 때문이다. 단, 이상값(outlier)은 작은 크기의 표본에 나타나지 않을수도 있다.

개별 속성의 값을 분석하는 또다른 방법은 적절한 요약 통계 지표를 (summary statistics) 사용하는 것이다. 분석의 목표에 따라 다양한 통계 지표가 존재하는데, 예컨데 데이터의 중심을 알기 위해서는 평균(mean) 및 중앙값(median), 최빈값(mode) 등을 [3] 사용할 수 있고, 데이터의 분산도를 알기 위해서는 범위(range), 분산(variance) 등을 사용할 수 있다. 또한, 데이터의 분포가 한쪽으로 기울었는지를 나타내는 통계 지표도 (skewness) 존재한다.

이런 통계 지표를 사용할 때에는 데이터의 특성에 주의해야 한다. 예컨데 평균에는 집합 내 모든 데이터의 값이 반영되기 때문에 이상값(outlier)이 존재하는 경우 값이 영향을 받지만, 중앙값에는 가운데 위치한 값 하나가 사용되기 때문에 이상값의 존재에도 대표성이 있는 결과를 얻을 수 있다. 예컨데, 회사에서 직원들의 평균 연봉을 구하면 중간값보다 훨씬 크게 나오는 경우가 많은데, 이는 몇몇 고액 연봉자들의 연봉이 전체 평균을 끌어올리기 때문이다.

개별 속성 분석에서 또한 중요한 부분은 시각화를 적절히 활용하는 것이다. 많은 수의 개별 데이터를 일일히 보는 것은 비효율적이고, 모든 데이터를 수치 하나로 요약하는 통계값에는 중요한 패턴이 감추어지는 경우가 많은데, 적절한 시각화를 통해 데이터에 나타나는 패턴을 한눈에 볼 수 있기 때문이다.

역시 데이터의 유형 및 분석의 초점에 따라 다양한 시각화 방법이 존재한다. 수치 및 카테고리 데이터의 경우 그 분포를 한눈에 볼 수 있는 히스토그램 등이 일반적이다. 또한 텍스트 데이터의 경우 단어의 밀도를 표현할 수 있는 워드 클라우드를, 시간 및 공간 데이터는 시계열차트 및 지도 등 해당 데이터를 가장 잘 표현할 수 있는 시각화 방식을 사용하면 된다. 아래 그림은 데이터 형태별로 적절한 시각화 방식을 보여준다.

이제 사례를 통해 개별 속성을 분석하는 요령을 알아보자. 여기서 사용할 데이터는 다양한 제조사 및 모델별로 1999년과 2008년형 자동차의 연비를 기록한 MPG라는 (MPG는 Mile Per Gallon의 약자로 미국에서 쓰는 자동차 연비의 단위다.)이름의 데이터셋이다. 이 데이터셋은 적당한 크기에 다양한 유형의 속성을 포함하고 있어서 탐험적 데이터 분석의 프로세스를 이해하기에 유용하다. MPG 데이터셋은 R에 기본적으로 포함되어 있으며, 이곳에 언급된 분석과 시각화를 수행하는 코드와 수행 결과를 참고하자.

본격적인 실습에 나서기 전에 분석의 목표를 생각해 보자. 우선 자동차의 연비를 다루는 데이터셋이니만큼 연비에 영향을 미치는 다양한 요인을 알아볼 수 있을 것이다. 또한 데이터에 연도와 제조사 및 자동차 모델 속성이 포함된 만큼 연도별로 연비가 어떻게 달라졌는지, 그리고 제조사와 모델별 연비는 어떤지 살펴볼 수 있을 것이다.

아래는 우리가 사용할 MPG 데이터의 다양한 속성값이다. 일단 눈으로 보면서 데이터에 익숙해지도록 하자. 데이터에 대한 설명을 보면 drv는 구동방식 (r: 후륜 / f: 후륜 / 4: 4륜), displ은 배기량, cyl은 실린더 개수를 나타내며, cty는 도시 주행시 연비를, hwy는 고속도로 주행시 연비를 나타낸다.

MAKER	CLASS	MODEL	YEAR	DISPL	CYL	DRV	CTY	HWY
audi	compact	a4	1999	1.8	4	f	18	29
audi	compact	a4 quattro	1999	1.8	4	4	18	26
audi	midsize	a6 quattro	1999	2.8	6	4	15	24
chevrolet	suv	c1500 suburban 2wd	2008	5.3	8	r	14	20
chevrolet	2seater	corvette	1999	5.7	8	r	16	26
chevrolet	suv	k1500 tahoe 4wd	2008	5.3	8	4	14	19

개별 데이터를 살펴본 후에는 데이터를 요약하는 다양한 통계값(statistics)을 계산하여 살펴볼 수 있다. 이런 통계값은 데이터의 유형에 따라 달라진다. 예컨대 연속된 수치에 대해서는 평균 및 중간값 등을, 그리고 범주 및 순서를 나타내는 속성에 대해서는 최빈값을 계산할 수 있을 것이다. 아래 표는 MPG 데이터의 다양한 속성들의 통계값이다.

데이터에 적절한 시각화를 적용하여 전체 값의 분포를 한눈에 볼 수 있다. 아래는 MPG 데이터셋의 고속도로 연비(hwy)를 확률 밀도 그래프, 히스토그램, 그리고 점플롯(dotplot)으로 표현한 것이다. 이 시각화 결과에 따르면 데이터에 크게 두가지 그룹의 차량이 (세단과 SUV) 있다는 사실이 드러난다. 이중 히스토그램과 점플롯에서는 개별 데이터의 분포를 좀더 뚜렷히 볼 수 있다.

속성간의 관계 분석하기

개별 속성의 분포를 확인한 후에는 속성간의 관계를 살펴보자. 이 과정의 목표는 서로 의미있는 상관관계를 갖는 속성의 조합을 찾아내는 것이다. 여기서부터 본격적으로 '탐험적' 분석이 시작되는데, 속성이 많은 데이터의 경우 속성의 조합에 따라 다양한 분석을 수행할 수 있기 때문이다. 모든 속성간의 관계를 다 보기 힘들 경우에는 주어진 문제 해결과 관련이 깊은 부분부터 시작해야 할 것이다.

이제 속성간의 관계를 분석하는 과정를 MPG 데이터를 통해 알아보도록 하자. 우선 아래 표는 수치형 속성간의 상관계수를 나타낸 것이다. 연비(hwy/cty)와 실린더 수 (cyl) 및 배기량(disp)이 높은 상관관계를 가지는 것을 볼 수 있다.

상관계수	DISPL	YEAR	CYL	CTY	HWY
displ	1.000	0.148	0.930	-0.799	-0.766

상관계수	DISPL	YEAR	CYL	CTY	HWY
year	0.148	1.000	0.122	-0.037	0.002
cyl	0.930	0.122	1.000	-0.806	-0.762
cty	-0.799	-0.037	-0.806	1.000	0.956
hwy	-0.766	0.002	-0.762	0.956	1.000

아래 플롯은 위 상관도 테이블을 그대로 시각화한 것이다. 오른쪽 위를 향하는 파란색 타원은 양의 상관관계를, 오른쪽 아래를 향하는 빨간색 타원은 음의 상관관계를, 흰색 원은 상관관계가 없음을 나타낸다.

하지만 상관계수가 데이터에 존재하는 모든 트랜드를 요약하는 것은 아니다. 같은 데이터에 해당하는 아래 스케터플롯을 보면 각 속성 쌍의 관계가 다양한 양상의 띈다는 것을 알 수 있다.

이렇게 모든 속성간의 관계를 상관도와 스케터플롯을 통해 관찰한 후에는 관심이 가는 개별 속성간의 관계를 자세히 살펴보아야 할 것이다. 두 속성간의 관계를 알아보는 방법은 대상 속성의 유형의 조합에 따라 달라진다. 우선 **두 카테고리형 속성간의 관계**를 생각해보자. 두 속성이 모두 제한된 카테고리의 값을 갖기 때문에, 이를 분석하는 데에는 각 카테고리의 조합에 속하는 항목의 수를 나타내는 교차 테이블이나 모자이크 플롯이 적절하다.

다음 표와 모자이크 플롯은 MPG 데이터셋에서 실린더의 수와 구동방식의 관계를 나타낸다. 후륜 방식은 대부분 대형차에, 그리고 전륜 방식은 소/중형차에 사용되는 것을 알 수 있다. 같은 데이터를 모자이크 플롯으로 보면 대부분의 차가 사륜구동 혹은 전륜구동을 띄고 있다는 사실, 그리고 실린더 수의 분포가 구동방식에 따라 다르다는 사실을 한눈에 알 수 있다.

구동방식	사륜(4)	전륜(F)	후륜(R)
4	23	58	0
5	0	4	0
6	32	43	4
8	48	1	21

다음으로 **카테고리형 속성과 수치형 속성간의 관계**를 분석하는 경우에는 제한된 종류의 값을 갖는 카테고리형 속성의 특성을 고려하여 각 카테고리 별로 수치값의 분포를 볼 수 있다. 여기에는 박스플롯이 널리 사용되는데, 박스플롯은 주어진 데이터의 25% 및 75%에 해당하는 백분위 값을 박스로 표시하고, 여기에 50%에 해당하는 중앙값을 가운데 굵은 선으로 표시한 플롯이다.

데이터를 대표하는 통계값을 한눈에 보기위해 박스플롯을 사용한다면, 개별 데이터 값을 관찰하기 위해서는 스케터플롯을 사용할 다. 단, 이 경우 같은 카테고리를 갖는 값들이 서로 겹치지 않도록 노이즈(jitter)를 더해준다.

아래 플롯은 연도와 고속도로 연비간의 관계를 박스플롯과 스케터플롯으로 보여준다. 아래 그림의 왼쪽의 박스플롯에는 1999년에 비해 2008년에 연비간의 관계가 뚜렷히 드러나지 않는다. 하지만 오른쪽의 스케터플롯을 보면 몇개의 이상값을 제외하고는 전반적인 연비가 더 높은 위치에 분포함을 알 수 있다.

마지막으로 **두 수치형 속성간의 관계**는 스케터플롯으로 확인하는 것이 가장 일반적이다. 아래 왼쪽의 스케터플롯에서는 배기량과 도시 연비의 상관관계를 보여준다. (상관계수:-0.799) 배기량과 도시 연비가 대체로 음의 상관관계를 갖지만, 어느정도 배기량이 높아지면 상관관계가 희박해진다. 이에 반해 도시 연비와 고속도로 연비에 관한 아래 오른쪽의 스케터플롯에서는 대부분의 값들이 좀 더 추세선 근처에 모여있는 것을 알 수 있다. (상관계수:0.956)

탐험적 분석에서는 두개 이상의 속성간의 관계를 동시에 보고 싶은 경우도 생긴다. 이럴 때는 기존의 스케터플롯이 유용한데, X와 Y의 위치에 두가지 속성의 관계를 표시하는 스케터플롯의 각 점에 색상, 크기, 레이블 등의 다양한 속성을 추가할 수 있기 때문이다.

다시 MPG 데이터셋을 살펴보면 아래 차트에는 고속도로 및 도시 주행시 연비를 XY축으로 하고, 여기에 배기량을 점의 크기에, 그리고 실린더의 개수를 점의 색상에 적용했다. 아래 그래프를 보면 크고 밝은 점이 (실린더 수가 많고 배기량이 높은 차량) 그래프 좌측 하단에 몰려있는 것을 볼 수 있다. 이처럼 스케터플롯을 사용해 네가지 속성의 관계를 동시에 확인할 수 있다.

마지막으로 산포도 그래프에 텍스트 레이블을 적용한 사례를 살펴보자. 아래 그래프는 점 대신 각 차량 모델의 연비를 그래프에 나타낸다. 이때 레이블의 색상으로 제조사를 구분할 수 있도록 하였다. 아래 그래프를 보면 연비가 낮은 대부분의 차량은 4륜 구동의 SUV 및 트럭임을 알 수 있다. 또한 높은 연비를 자랑하는 차량은 대부분 한국/일본산이거나 그리고 독일의 폭스바겐에서 만들어진 것들이다. 또한 아우디에서 만든 차량은 고속도로 연비에 비해 도시 연비가 좋지 않다는 것도 알 수 있다.

적절한 시각화의 중요성

지금까지 탐험적 데이터 분석의 프로세스를 알아보았다. 위의 다양한 사례에서 드러났지만, 여기서 다시 한번 강조하고 싶은 것은 나무와 숲을 동시에 볼 수 있게 해주는 시각화의 중요성이다. 특히 이런 관점에서 전체 추세와 함께 개별 데이터의 분포를 그대로 보여주는 스케터플롯이 특히 유용하다.

여기서는 몇가지 사례를 통해 개별 데이터의 분포를 확인하는 작업의 중요성을 알아보자. 우선 아래 그림은 같은 평균과 분산을 갖는 (그림 A) 두 데이터 집단의 실제 분포가 (그림 B~E) 얼마나 다를 수 있는지를 보여준다. 막대그래프로는 분간하기 힘든 두 집단간의 차이가 스케터플롯으로 명쾌하게 드러나는 것을 알 수 있다. [@Weissgerber2015]

비슷한 사례로 이번에는 두가지 수치 형태의 속성간의 관계를 시각화하는 경우를 생각해보자. 아래 그림은 각기 다른 데이터셋 4개를 산포도 플롯으로 표시한 것이다. 얼핏 각 데이터셋은 파란 선에 걸쳐있다는 점을 제외하고는 전혀 다른 것처럼 보인다.

하지만 각 데이터를 가지고 다음 표의 각종 통계값을 계산해보면 위 플롯의 데이터셋 4개는 모두 같은 평균, 분산, 그리고 상관계수를 가진다. 시각화 없이 통계값에만 의존했다면 이런 차이를 전혀 알 수 없었을 것이다.

통계 유형	값
x의 평균	9
x의 분산	11
y의 평균	7.5
y의 분산	4.122 혹은 4.127
x와 y의 상관계수	0.816

마지막으로 아래 인포그래픽은 각기 다른 종류의 시각화 기법을 분석의 목표 및 데이터의 특성에 따라 어떻게 적용할 것인지를 요약해서 설명한다. 우선 분석의 목표를 수치 비교 / 분포 시각화 / 관계 시각화 / 구성 시각화 중 하나로 결정하고, 데이터의 분포도의 특성에 따라 세부적인 시각화 방법을 고를 수 있다. (출처: http://www.extremepresentation.com/) [@ExtremePresentation2011]

맺음말

지금까지 탐험적 데이터 분석의 절차를 개별 속성 분석과 속성간 관계 분석의 순서로 알아보았다. 또한 데이터 시각화의 중요성과 유의사항을 살펴보았다. 탐험적 분석을 성공적으로 마쳤다면 주어진 데이터를 속속들이 이해함과 동시에, 데이터에 대한 다양한 가설을 세울 수 있을 것이다. 이런 과정은 현상에 대한 더 나은 이해와 함께, 주어진 데이터를 가지고 다양한 예측 모델을 만들거나 (기계학습), 신뢰성있는 결론을 내리는데 (통계적 추론) 기반이 된다.

데이터 분석의 올바른 태도에 대해 생각해보는 것으로 이 장을 마무리할까 한다. 앞서 문제정의 단계에서 데이터를 모으기 전에 주어진 문제의 목표와 범위, 그리고 예상되는 결론(가설)을 최대한 고민해야 한다고 이야기했다. 하지만 **데이터를 막상 받아든 분석가는 백지 상태에서 데이터를 보기 위해 노력해야 한다**. 데이터를 지나친 편향된 시각에서 본다면 데이터를 있는 그대로 해석하기보다는 자신의 기대에 맞는 부분에만 집중하는 오류를 범할 수 있기 때문이다.

이렇게 문제 해결의 단계에 따라 관점을 계속 바꿀 수 있어야 한다는 점이 데이터 과학의 어려움이지만, 예술 및 과학을 포함한 모든 창조적인 작업은 자신의 창조물에 대한 열정을 가짐에 동시에 (예: 과학자의 가설) 이를 끊임없이 객관적이고 비판적인 시각에서 볼 수 있는 냉철함을 요구한다. 하지만 이런 노력 끝에 얻어지는 결과물을 주관적 만족감과 객관적인 가치를 동시에 충족시킬 수 있을 테니 그만한 가치가 있다고나 할까?

(본 글의 초고에 피드백을 보내주신 권정민 / 이성훈 님을 비롯한 독자 분들께 감사의 말씀을 전합니다.)

1. 나탈리 골드버그의 책 '뼈속까지 내려가서 써라'에는 글쓰기가 지도없이 떠나는 여행이라는 표현이 등장한다. ↩
2. 나무와 숲을 동시에 볼 수 있게 해주는 스케터플롯은 이런 관점에서 유용한 도구다. ↩
3. 이들 지표에 대한 자세한 설명은 위키피디아를 참조하도록 하자. ↩

Share this:

Like this:

Related

데이터과학 프로세스(3): 데이터 준비하기
July 14, 2015
In "데이터과학 프로세스"

데이터과학 프로세스(1): 데이터 문제 정의하기
June 1, 2015
In "데이터과학 프로세스"

'데이터' 과학 개념잡기: 빅데이터의 허상과 스몰데이터의 가치
March 18, 2015
In "데이터과학 개념잡기"

This entry was posted in 데이터과학 프로세스 on July 25, 2015 [http://www.hellodatascience.com/?p=323] by lifidea.