滴滴出行大数据预测体系之“猜您要去”目的地预测系统

分享人：张凌宇
2016.10.22

滴滴
滴滴一下 美好出行

业务场景


1
分钟
在这里上车
定阜街
龙头井街
前海西街
兴化胡同
东北旺数字山谷B区 [猜您要去]
个人支付
拼车
7.2 元起
不拼车
约 9.8 元

猜您要去 用户反响

一滴水 👍 1

佩服佩服啊太准确了😱

华莹 👍 2

炫酷

小黑 👍 3

功能吊爆

晶晶 👍 12

哇塞，猜得好准！

小琼【中国康嘉奇】 👍 1

太牛□ 逼了吧！这个好😁😁😁

这个功能有什么用呢

- 降低发单的输入成本
  - 一键发单，告别手机打字烦恼
- 惊艳用户，彰显滴滴的人工智能科技
  - 以90+%的准确率，预测30+%的出行
- 预测出行流向，更好的规划交通运力
  - 一大波人群将要去往xxx

目录

这是个什么问题

- 产品经理：咱们有个“猜你去哪”的需求
- 研发工程师：猜啥？
- 产品经理：猜目的地
- 研发工程师：啥场景下猜？
- 产品经理：猜当前时间、当前地点出发的订单的目的地
- 研发工程师：咋猜？
- 产品经理：根据用户出行历史记录猜

问题定义：通过用户出行历史，预测当前地点、当前时间下的出行目的地

模型抽象

问题定义：通过用户出行历史，预测当前地点、当前时间下的出行目的地

T：当前时间，S：当前位置，x：预测目的地，$\{x_k|k \in [1,n]\}$：用户历史目的地 集合。

则被预测的目的地 x_i满足以下条件：

$$\exists x_i,$$
$$P(X = x_i|t = T, s = S) \gg \sum_{k=1,k\neq i}^{n} P(X = x_k|t = T, s = S)$$

问题转化为：对$\{x_k|k \in [1,n]\}$集合，计算 $P(X = x_k|t = T, s = S)$。

目录

特征分析

对$\{x_k|k \in [1,n]\}$集合，计算 $P(X = x_k|t = T, s = S)$

在目标变量$P(x|t,s)$中，t和s均为复合变量，其中，

t包括日期（date）和时刻（time）

如 “2016-08-23 18:10:10”；

s包括地址名称（address）和经纬度（(lat, lng)）

如“天安门；39.341231,116.23521”

目标变量变为：$P(x|date, time, address, (lat, lng))$

特征分析

$\{date,\ time,\ address,\ (lat, lng)\}$ 四个特征变量中，包含了如下类型：

离散型变量：address

连续型变量：lat、lng、date

周期型变量：time

二维联合变量：(lat、lng)

特征选择

快速搭建模型：选择最主要特征，忽略次要特征

模型持续优化：持续增加特征，挖掘特征间的关系

度量变量间相关性的指标：

皮尔逊系数：

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

互信息：

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

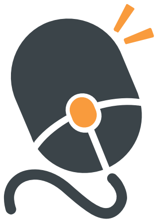
各特征与目标变量间的互信息

选取最近90天的订单，按用户分组后，计算每个用户下的互信息

目的地 ：x，**出发地**：f，**出发时刻**：t，**出发日期属性（周末or工作日）**：d

I(x, f)	I(x, d)	I(x, t)
0.82	0.43	**1.26**

结论：单维度特征下，出发时刻是预测目的地 的最好的特征



连续特征下的贝叶斯估计

问题简化为：求解$P(X = x_i|T = t)$的概率模型。

根据**贝叶斯公式**和**全概率公式**

$$P(X = x_i|T = t) = \frac{P(T = t|X = x_i) * P(X = x_i)}{P(T = t)}$$

$$P(T = t) = \sum [P(T = t|X = x_i) * P(X = x_i)]$$

$$P(X = x_i|T = t) = \frac{P(T = t|X = x_i) * P(X = x_i)}{\sum [P(T = t|X = x_i) * P(X = x_i)]}$$

问题进一步转化为求解$P(T|X)$的概率分布

目录

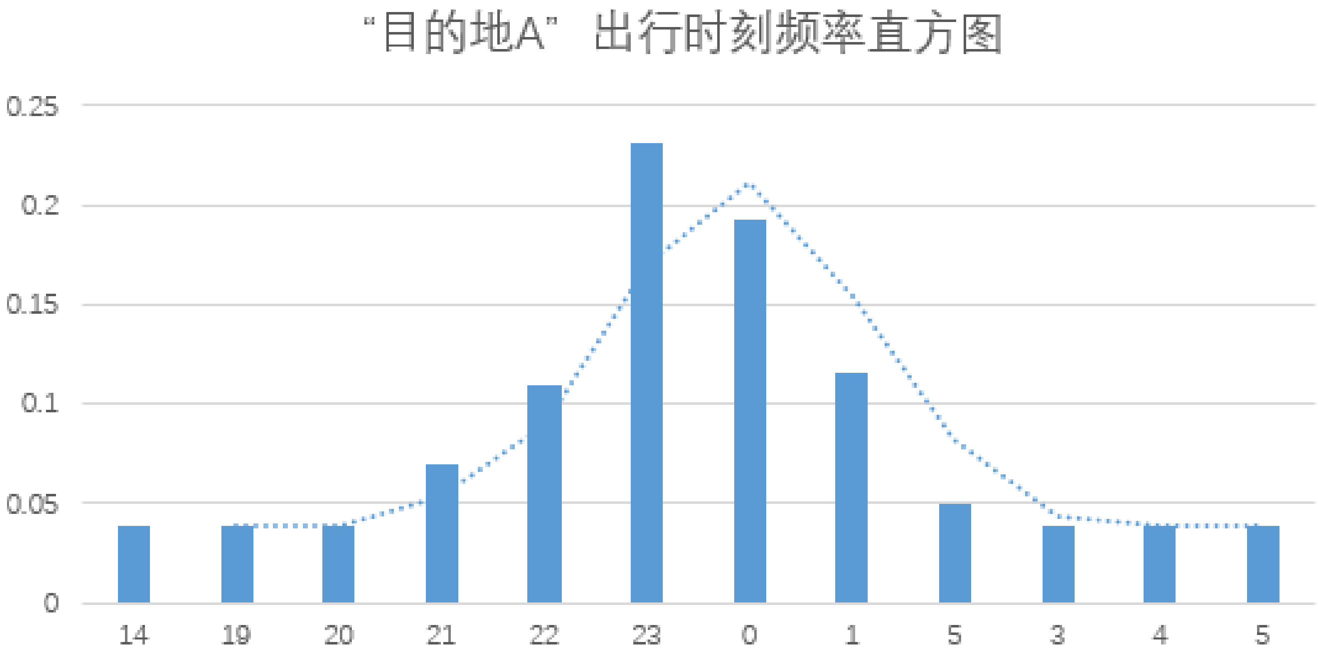
一个用户的出行case

出发时间	目的地
2016/1/12 23:52	目的地 A
2016/1/12 21:32	目的地 B
2016/1/7 21:12	目的地 A
2015/12/29 23:06	目的地 G
2015/12/29 11:56	目的地 B
2015/12/28 21:17	目的地 H
2015/12/28 11:44	目的地 B
2015/12/21 14:54	目的地 A
2015/12/20 13:01	目的地 B
2015/12/20 11:25	目的地 J
2015/12/19 23:18	目的地 A
2015/12/19 19:36	目的地 P
2015/12/19 1:14	目的地 A
2015/12/18 20:03	目的地 K
2015/12/18 7:46	目的地 B
2015/12/18 2:02	目的地 A
2015/12/17 23:12	目的地 J
2015/12/17 11:28	目的地 B
2015/12/14 23:59	目的地 A
2015/12/12 22:00	目的地 O
2015/12/12 19:29	目的地 A
2015/12/12 17:53	目的地 Q
2015/12/12 12:29	目的地 L
2015/12/11 23:25	目的地 A

一个用户的出行case

目的地	所有发单时刻
目的地 A	0,0,0,0,1,1,1,2,2,2,2,2,3,4,5,14,19,20,21,21,21,22,22,23,23,23,23,23,23,23
目的地 B	7,7,8,8,8,9,9,10,11,11,11,11,11,12,12,13,13,13,14,16,16,17,21,21,21
目的地 C	16
目的地 D	17,23
目的地 E	16
目的地 F	8
目的地 G	23
目的地 H	21
目的地 I	14
目的地 J	11,23
目的地 K	20
目的地 L	21
目的地 M	12
目的地 N	10
目的地 O	22
目的地 P	17
目的地 Q	12

一个用户的出行case



正态分布及对应的参数估计

对2000多个case进行类似上面的分析，基本符合**正态分布**

因此，我们采用正态分布进行近似估计

$$P(T|X) \sim N(\mu, \sigma)$$

所以接下来的问题就是：

如何估计这个分布的μ和σ

模型的关键部分：时刻的均值和方差的估计

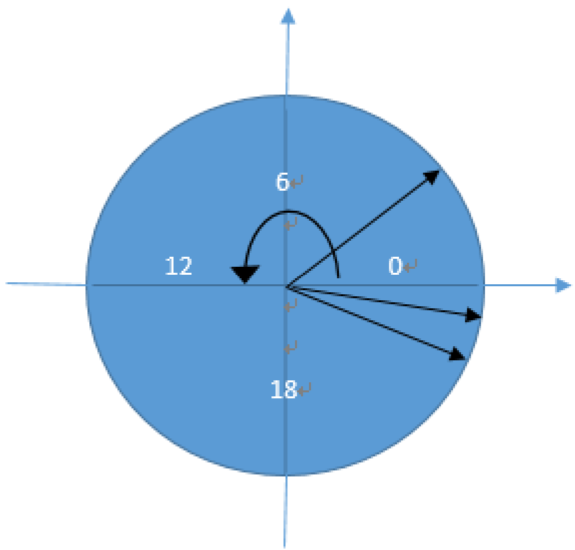
- 8点、9点、10点，均值是9点；
- 23点、0点、1点，均值是0点；
- 3点、22点、23点，均值是？
- 3点、15点、21点，均值是？

时刻均值和方差的估计——向量法

3点、22点、23点，平均时刻是多少？

这里将每个时刻转化为向量表示法，如下如所示：

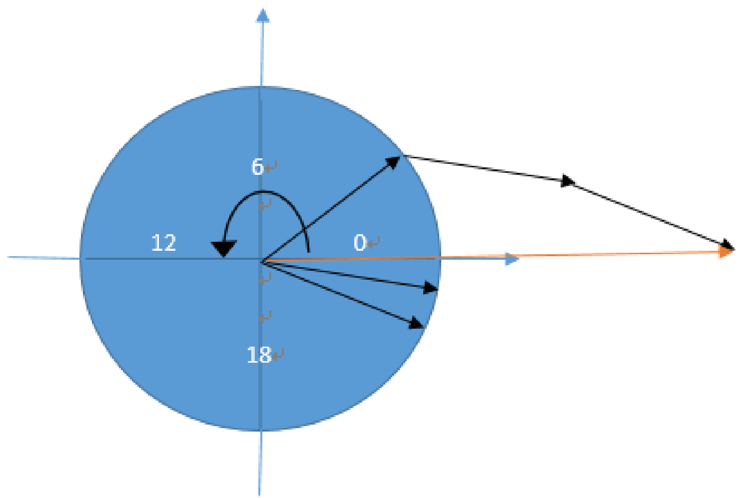
圆盘表示时钟表盘，两个坐标轴分别为x、y轴，图上的三个向量分别表示的时刻是：3点、22点、23点，圆弧箭头表示时钟的方向。



时刻均值和方差的估计——向量法

3点、22点、23点，平均时刻是多少？

将这三个向量加和，和向量落在的表盘上的位置即是平均时刻，如下图：



3点、22点、23点三个时刻的平均时刻是0点，和向量落在的表盘上的位置也恰好是0点。

时刻均值和方差的估计——向量法

Step1： 第i个时刻x_i的向量表示

$$(\cos\theta_i, \sin\theta_i)$$

$$\theta_i = 2\pi * \frac{x_i}{24}$$

Step2： 计算n个时刻对应的向量的和向量

$$\left(\sum_{i=1}^{n}\cos\theta_i, \sum_{i=1}^{n}\sin\theta_i\right)$$

Step3： 计算和向量与x轴的夹角：

$$\theta_t = \cos^{-1}\frac{\sum_{i=1}^{n}\cos\theta_i}{\sqrt{\left(\sum_{i=1}^{n}\cos\theta_i\right)^2 + \left(\sum_{i=1}^{n}\sin\theta_i\right)^2}}$$

Step4： 将θ_t转换为对应的时刻：

$$\mu = 24 * \frac{\theta_t}{2\pi}$$

Step5： 对应的方差计算公式：

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(-\big|\,|x_i - \mu| - 12\,\big| + 12\right)^2$$

时刻均值和方差的估计——向量法

时刻	均值（向量法）	均值（理论上）
00:00:00 00:00:00 03:00:00	00:58:33	01:00:00
00:00:00 12:00:00	无解	06:00:00 or 18:00:00

问题来了：向量法只能得到近似解，且在边界情况下无解，肿么办？

时刻均值和方差的估计——拉格朗日法

算数平均值的一个重要性质：

算数平均值与所有观测样本的距离平方和最小

就是下面这个优化问题的解：

$$t.g.:\min\sum_{i=1}^{n}(X_i-\overline{X})^2$$

时刻均值和方差的估计——拉格朗日法

求解过程

$$\mathrm{L}(\overline{X}) = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\frac{dL}{d\overline{X}} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)$$

$$\frac{dL}{d\overline{X}} = 0 \quad \Rightarrow \quad \boldsymbol{\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

时刻均值和方差的估计——拉格朗日法

按照上面的逻辑，

平均时刻也可以认为是与所有时刻距离平方和最小的那个时刻

即下面这个二次优化问题解：

$$\begin{cases} t.g.: min \sum_{i=1}^{n} [distance(X_i, \overline{X})]^2 \\ X_i \in [0,24) \\ s.t.\ \ \overline{X} \in [0,24) \end{cases}$$

这里又引入了一个新的概念：两个时刻的距离。

时刻均值和方差的估计——拉格朗日法

$distance(T_1, T_2)$:

- 首先，该距离不能是负值，即 $distance(T_1, T_2) \geq 0$;
- 其次，该距离不能超过12，即 $distance(T_1, T_2) \leq 12$;

$$distance(T_1, T_2) = \begin{cases} |T_1 - T_2|, & if\ |T_1 - T_2| \leq 12 \\ 24 - |T_1 - T_2|, & if\ |T_1 - T_2| > 12 \end{cases}$$

时刻均值和方差的估计——拉格朗日法

$$distance(T_1, T_2)$$

$$= \begin{cases} |T_1 - T_2|, & if \ |T_1 - T_2| \le 12 \\ 24 - |T_1 - T_2|, & if \ |T_1 - T_2| > 12 \end{cases}$$

$$\Rightarrow \quad distance(T_1, T_2) = -\big|\, |T_1 - T_2| - 12 \,\big| + 12$$

时刻均值和方差的估计——拉格朗日法

最后，得到：

$$\begin{cases} t.g.: G \\ G = min\sum_{i=1}^{n}\left(-\left|\,|X_i - \overline{X}| - 12\,\right| + 12\right)^2 \\ X_i \in [0, 24) \\ s.t.\ \overline{X} \in [0, 24) \end{cases}$$

解这个优化问题，得到时刻的均值和方差

$$\begin{cases} \mu = \overline{X} \\ \sigma^2 = \dfrac{1}{n}G \end{cases}$$

向量法和拉格朗日法对比

向量法：

简洁，容易理解；

近似解，边界条件下无解；

拉格朗日法：

精确解；

直观上不好理解，求解算法略复杂；

循环正态分布

上面的假设，用正态分布去估计出发时刻的分布，

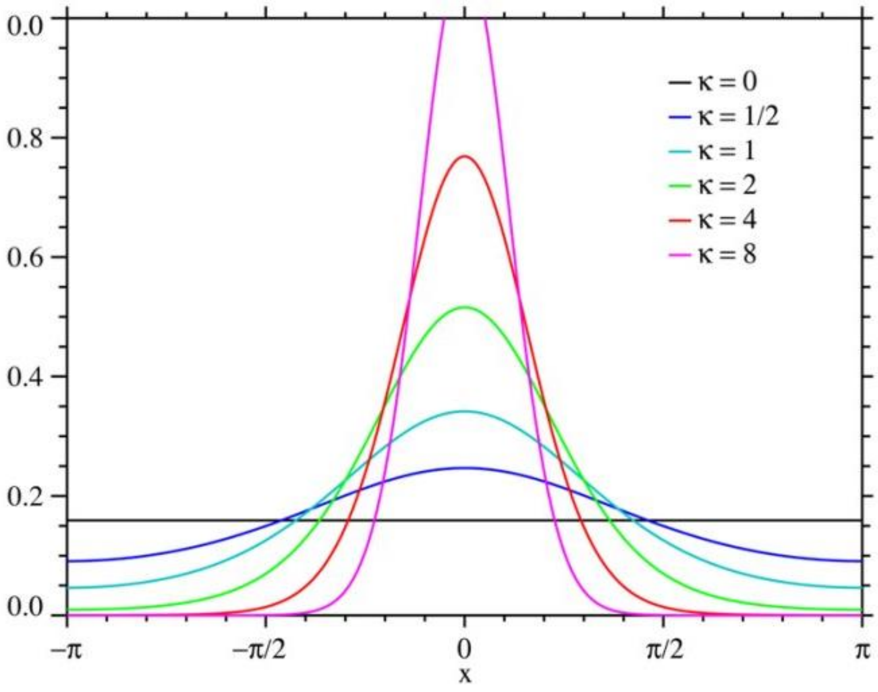
$$P(T|X) \sim N(\mu, \sigma)$$

实际上，
- 正态分布自变量的分布是整个数轴$(-\infty, +\infty)$
- 而出发时间T的分布是[0,24]，并且具有周期循环性。

在高级数理统计中，这种分布叫做循环正态分布
——冯·米塞斯分布。

其概率密度函数为： $f(x|\mu,\kappa) = \dfrac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$

冯•米塞斯分布的图像



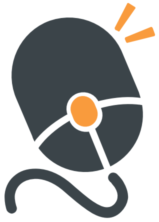
算法流程

Step1：根据该用户的订单历史，计算每个目的地 的发单时刻集合的μ和σ；

Step2：根据当前时间，计算每个目的地 的$P(T|X_i)$和频率$P(X_i)$；

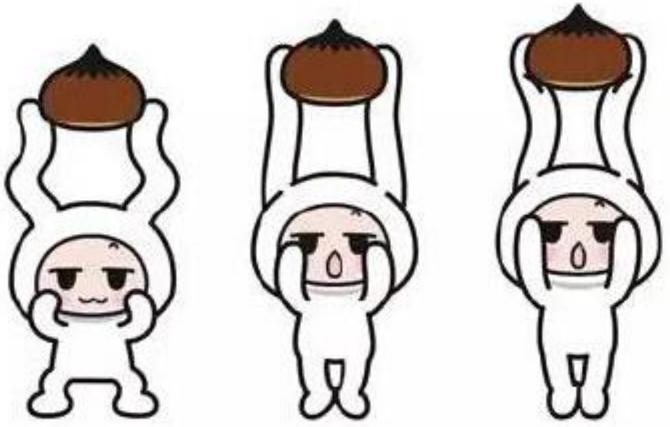
Setp3：计算每个目的地 的概率

$$\mathrm{P}(X_i|T) = \frac{P(T|X_i) * P(X_i)}{\sum[P(T|X_i) * P(X_i)]}$$



Step4：确定支持度阈值s和概率阈值p，对满足阈值的地址作为预测结果。

举个栗子



举个栗子

目的地	时间分布	分布指标
目的地 A	8.7,9.7,9.9,9.9,9.9,10,10.1,10.1,16	e=10.5, d=2, f=0.47
目的地 B	18,18.2,18.9,19,19.3,20.5,21.1	e=19.3, d=0.96, f=0.37
目的地 C	19,20	e=19.5, d=0.25, f=0.05
目的地 D	18,20	e=19, d=0.33, f=0.05
目的地 E	22,23	e=22.5, d=0.25, f=0.05

举个栗子

假设当前时刻 T=9 点

$P(T=9|X=\text{目的地 A}) = 0.3$

$P(T=9|X=\text{目的地 B}) = 0.05$

$P(T=9|X=\text{目的地 C}) = 0.02$

$P(T=9|X=\text{目的地 D}) = 0.02$

$P(T=9|X=\text{目的地 E}) = 0.01$

$P(X=\text{目的地 A}) = 0.47$

$P(X=\text{目的地 B}) = 0.37$

$P(X=\text{目的地 C}) = 0.05$

$P(X=\text{目的地 D}) = 0.05$

$P(X=\text{目的地 E}) = 0.05$

举个栗子

经过贝叶斯转化，最终得到各个目的地 的概率如下：

$P(X=\text{目的地 A}|\ T=9\text{点}) = 0.98125$

$P(X=\text{目的地 B}|\ T=9\text{点}) = 0.015625$

$P(X=\text{目的地 C}|\ T=9\text{点}) = 0.00625$

$P(X=\text{目的地 D}|\ T=9\text{点}) = 0.00625$

$P(X=\text{目的地 E}|\ T=9\text{点}) = 0.003125$

假设我们将概率阈值定为 0.9，

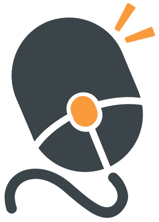
“目的地 A” 将出现在目的地框中。

目录

出发地经纬度建模

某乘客某一目的地 的出发地经纬度list

目的地 （D)	出发地经纬度（FL）
目的地 A	116.363868,39.915524;116.3638,39.9229; 116.341193,39.922947;116.34133,39.922956; 116.341614,39.922952;116.341514,39.922947; 116.341308,39.922947;116.341248,39.922764; 116.341161,39.922943;116.341056,39.922947; ………

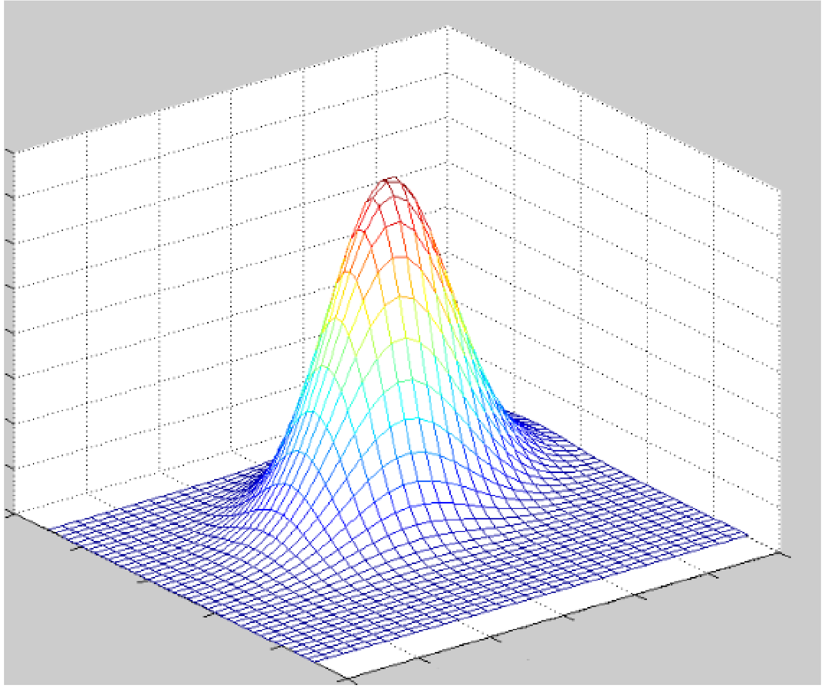


出发地经纬度建模

每个经纬度的频次和距离分布。

出发地经纬度（FL）	偏离中心点距离（米）	频次C
39.912,116.473	10279	1
39.914,116.474	9902	1
39.963,116.49	2220	1
39.988,116.492	4303	1
39.924,116.503	6920	3
39.965,116.5	506	4
39.964,116.503	168	9
39.965,116.503	0	24
39.966,116.503	168	16
39.964,116.504	238	2
39.965,116.504	168	1
39.966,116.505	377	1
39.958,116.507	1360	1
39.919,116.522	8400	1

出发地经纬度建模



出发地经纬度建模

所以，提出假设：

同一用户同一目的地 的出发地经纬度（X, Y）服从参数为

$(\mu1, \mu2, \sigma_1, \sigma_2, \rho)$二维正态分布，即$(X, Y) \sim N(\mu1, \mu2, \sigma_1^2, \sigma_2^2, \rho)$

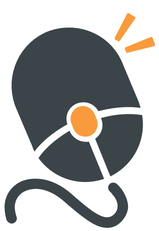
其密度函数为：

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu1)(y-\mu2)}{\sigma_1\sigma_2} + \frac{(y-\mu2)^2}{\sigma_2^2}\right]\right\}$$

出发地经纬度建模

最后，按照上面的模型，利用用户的出发地经纬度，对目的地 进行概率预测。

出发地经纬度（FL）	目的地（D）	频次（C）	概率 P(D|FL)
39.918560; 116.364716	目的地 A	9	0.81
	目的地 B	41	0.14
	目的地 C	3	0.05
	目的地 D	10	0
	目的地 E	1	0
	目的地 F	2	0
	目的地 G	1	0
	目的地 H	3	0



出发地与出发时间联合分布模型

经过前面的研究结论，我们已经知道：

1、$P(T|D) \sim N(\mu, \sigma)$

2、$P(Flat, Flng|D) \sim N(\mu 1, \mu 2, \sigma_1^2, \sigma_2^2, \rho)$

所以：$P(Flat, Flng, T|D) \sim N_3(\mu, C)$，三元正态分布。

出发地与出发时间联合分布模型

做一些变量代换，

$$
\begin{cases}
X = (T, Flat, Flng)^T \\
\mu = (E\{T\}, E\{Flat\}, E\{Flng\})^T \\
C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{12} & c_{22} & c_{23} \\ c_{13} & c_{23} & c_{33} \end{bmatrix} \\
c_{ij} = Cov(x_i, x_j) = E\{[x_i - \mu_i][x_j - \mu_j]\} \\
x_1 = T, x_3 = Flat, x_3 = Flng
\end{cases}
$$

$$
P(Flat, Flng, \mathrm{T}|D)
$$

$$
= P(X|D) = \frac{1}{(2\pi)^{3/2}} \frac{1}{\sqrt{\det C}} \exp\{-\frac{1}{2}(X - \mu)^T C^{-1} (\mathrm{X} - \mathrm{\mu})\}
$$

出发地与出发时间联合分布模型

以某个用户的某个目的地 为例，

先计算$(Flat, Flng, T)$三个变量的期望μ和协方差矩阵C,以及$\det C$和C^{-1}

$$
\begin{cases}
\mu = \begin{bmatrix} 20.35 \\ 40.028 \\ 116.534 \end{bmatrix} \\
C = \begin{bmatrix} 1.43414141e+01 & -5.29556407e-02 & -5.22035906e-02 \\ -5.29556407e-02 & 6.63847604e-03 & 6.08965387e-03 \\ -5.22035906\mathrm{e}-02 & 6.08965387\mathrm{e}-03 & 9.09750208\mathrm{e}-03 \end{bmatrix} \\
\det C = |C| = 0.000324359553166 \\
C^{-1} = \begin{bmatrix} 7.18636621\mathrm{e}-02 & 5.05187077\mathrm{e}-01 & 7.42101242\mathrm{e}-02 \\ 5.05187077\mathrm{e}-01 & 3.93840196\mathrm{e}+02 & -2.60728481\mathrm{e}+02 \\ 7.42101242\mathrm{e}-02 & -2.60728481\mathrm{e}+02 & 2.84871629\mathrm{e}+02 \end{bmatrix}
\end{cases}
$$

出发地与出发时间联合分布模型

带入

$$P(Flat,Flng,\mathrm{T}|D)=P(X|D)=\frac{1}{(2\pi)^{3/2}}\frac{1}{\sqrt{|C|}}\exp\{-\frac{1}{2}(X-\mu)^{T}C^{-1}(\mathrm{X}-\mathrm{\mu})\}$$

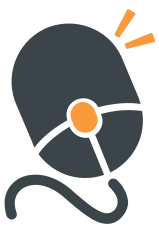
并展开，得到

$$\begin{cases} P(Flat,Flng,\mathrm{T}|D)=P(X|D)=\frac{1}{(2\pi)^{3/2}}\frac{1}{\sqrt{0.000324}}\exp\{\sum_{i,j=1}^{3}c_{ij}^{-1}(x_i-\mu_i)(x_j-\mu_j)\} \\ c_{ij}^{-1}=\frac{C_{ij}^{*}}{|C|} \\ C_{ij}^{*}\text{ 为 C的代数余子式} \end{cases}$$

出发地与出发时间联合分布模型

最后，按照上面的模型，利用用户的出发地经纬度、出发时刻，对目的地 进行概率预测。

出发地经纬度（FL） 时刻T	目的地 （D)	频次 （C)	概率 P(D | FL,T)
39.918560;116.364716 8:29:54	目的地 A	9	0.97
	目的地 B	41	0.03
	目的地 C	3	0
	目的地 D	10	0
	目的地 E	1	0
	目的地 F	2	0
	目的地 G	1	0
	目的地 H	3	0



高维特征下的逻辑回模型

多元正态分布模型：

$$P(X|D)=\frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{|C|}}\exp\{-\frac{1}{2}(X-\mu)^T C^{-1}(\mathrm{X}-\mathrm{\mu})\}$$

我们假设D是某一特定目的地，Y是目的地 取值变量，值域是个人历史目的地 列表，则

$$\begin{cases} P(X|Y=D)=\frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{|C|}}\exp\{-\frac{1}{2}(X-\mu)^T\ C^{-1}\ (\mathrm{X}-\mathrm{\mu})\} \\ P(X|Y\neq D)=\frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{|C'|}}\exp\{-\frac{1}{2}(X-\mu')^T\ C'^{-1}\ (\mathrm{X}-\mu') \end{cases}$$

$$\begin{cases} \mu\text{是}\mathrm{Y}=\mathrm{D}\text{条件下的}\mathrm{X}\text{的期望向量} \\ \mu'\text{是}\mathrm{Y}\neq\mathrm{D}\text{条件下的}\mathrm{X}\text{的期望向量} \\ C\text{是}\mathrm{Y}=\mathrm{D}\text{条件下的}\mathrm{X}\text{的协方差矩阵} \\ C'\text{是}\mathrm{Y}\neq\mathrm{D}\text{条件下的}\mathrm{X}\text{的协方差矩阵} \end{cases}$$

高维特征下的逻辑回模型

对P(Y=D|X)进行贝叶斯变换，得到

$$
\begin{aligned}
&P(Y = D|X) \\
&= \frac{\left(\int P(X|Y=D)dx\right) P(Y=D)}{\left(\int P(X|Y=D)dx\right) P(Y=D)+\left(\int P(X|Y\neq D)dx\right) P(Y\neq D)} \\
&\approx \frac{P(X|Y=D)\, P(Y=D)}{P(X|Y=D)\, P(Y=D)+P(X|Y\neq D)\, P(Y\neq D)} \\
&= \frac{1}{1+\frac{1-P(Y=D)}{P(Y=D)}\times\frac{P(X|Y\neq D)}{P(X|Y=D)}}
\end{aligned}
$$

高维特征下的逻辑回模型

下面处理 $\frac{P(X|Y\neq D)}{P(X|Y=D)}$

$$\frac{P(X|Y\neq D)}{P(X|Y=D)}$$

$$=\frac{\frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{|C|}}\exp\{-\frac{1}{2}(X-\mu)^{T}\,C^{-1}\,(\mathrm{X}-\mu)\}}{\frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{|C'|}}\exp\{-\frac{1}{2}(X-\mu')^{T}\,C'^{-1}\,(\mathrm{X}-\mu')\}}$$

$$=\sqrt{\frac{|C'|}{|C|}}\exp\{-\frac{1}{2}[(X-\mu)^{T}\,C^{-1}\,(\mathrm{X}-\mu)-(X-\mu')^{T}\,C'^{-1}\,(\mathrm{X}-\mu')]\}$$

高维特征下的逻辑回模型

先看$(X-\mu)^T C^{-1} (\mathrm{X}-\mathrm{\mu})$，展开：

$$(X^T-\mu^T)\, C^{-1}(X-\mu)$$

$$= X^T C^{-1}X - X^T C^{-1}\mathrm{\mu} - \mu^T C^{-1}X + \mu^T C^{-1}\mathrm{\mu}$$

$$= X^T C^{-1}X - 2\mu^T C^{-1}X + \mu^T C^{-1}\mathrm{\mu}$$

上面用到了$(X^T C^{-1}\mathrm{\mu})^T = X^T C^{-1}\mathrm{\mu}$和$(C^{-1})^T = C^{-1}$

于是，

$$[(X-\mu)^T C^{-1} (\mathrm{X}-\mathrm{\mu}) - (X-\mu')^T C'^{-1} (\mathrm{X}-\mu')]$$

$$= X^T(C^{-1} - C'^{-1})X - 2(\mu^T C^{-1} - \mu'^T C'^{-1})X + (\mu^T C^{-1}\mathrm{\mu} - \mu'^T C'^{-1}\mu')$$

高维特征下的逻辑回模型

带入上面的式子，得到：

$$P(Y=D|X)$$

$$=\frac{1}{1+\frac{1-P(Y=D)}{P(Y=D)}\times\frac{P(X|Y\neq D)}{P(X|Y=D)}}$$

$$=\frac{1}{1+\frac{1-P(Y=D)}{P(Y=D)}\times\sqrt{\frac{|C'|}{|C|}}\times\exp\{-\frac{1}{2}[(X-\mu)^T C^{-1}(X-\mu)-(X-\mu')^T C'^{-1}(X-\mu')]\}}$$

$$=\frac{1}{1+\frac{1-P(Y=D)}{P(Y=D)}\times\sqrt{\frac{|C'|}{|C|}}\times\exp\{-\frac{1}{2}[X^T(C^{-1}-C'^{-1})X-2(\mu^T C^{-1}-\mu'^T C'^{-1})X+(\mu^T C^{-1}\mu-\mu'^T C'^{-1}\mu')]\}}$$

高维特征下的逻辑回模型

做了一些变量代换：

$$
\begin{cases}
k = \ln\left(\frac{1-P(Y=D)}{P(Y=D)} \times \sqrt{\frac{|C'|}{|C|}}\right) \\
A = -\frac{1}{2}\left(C^{-1} - C'^{-1}\right) \\
\theta^T = \left(\mu^T C^{-1} - \mu'^T C'^{-1}\right) \\
b = -\frac{1}{2}\left(\mu^T C^{-1}\mu - \mu'^T C'^{-1}\mu'\right) + k
\end{cases}
$$

最后得到：

$$P(Y = D|X) = \frac{1}{1+\exp[X^T A X + \theta^T X + b]}$$

高维特征下的逻辑回模型

几点注意：

- 1、对比标准的逻辑回归，要加上变量的二次项和交叉项；
- 2、实际的数据，未必符合正态分布或规律性很强的分布；

特征工程——特征筛选

选择与目标变量相关性较高的特征

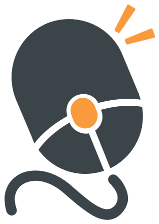
- 出发时间
- 出发地
- 用户信息

特征工程——特征拆分

利用业务常识，将选取的特征拆分成多个更细粒度的子特征

$$T\text{（时间特征）} => \begin{cases} \text{时刻，8点、9点} \\ \text{时段，早、中、晚、夜} \\ \text{星期} \\ \text{周中}\,or\,\text{周末} \\ \dots\dots \end{cases}$$

$$\boldsymbol{F}\text{（出发地特征）} => \begin{cases} \text{城市} \\ \text{行政区域} \\ \text{商圈} \\ \text{周围打车情况} \\ \text{经纬度} \\ \dots\dots \end{cases}$$



特征工程——特征挖掘

对用户行为进行分析，挖掘用户行为背后的隐含特征

$$\boldsymbol{P} \; => \begin{cases} \text{公司地址（用户主动填写）} \\ \text{家庭住址（用户主动填写）} \\ \text{通勤习惯} \\ \text{常住地} \\ \text{用户画像} \\ \dots\dots \end{cases}$$

特征工程——特征提取和离散化

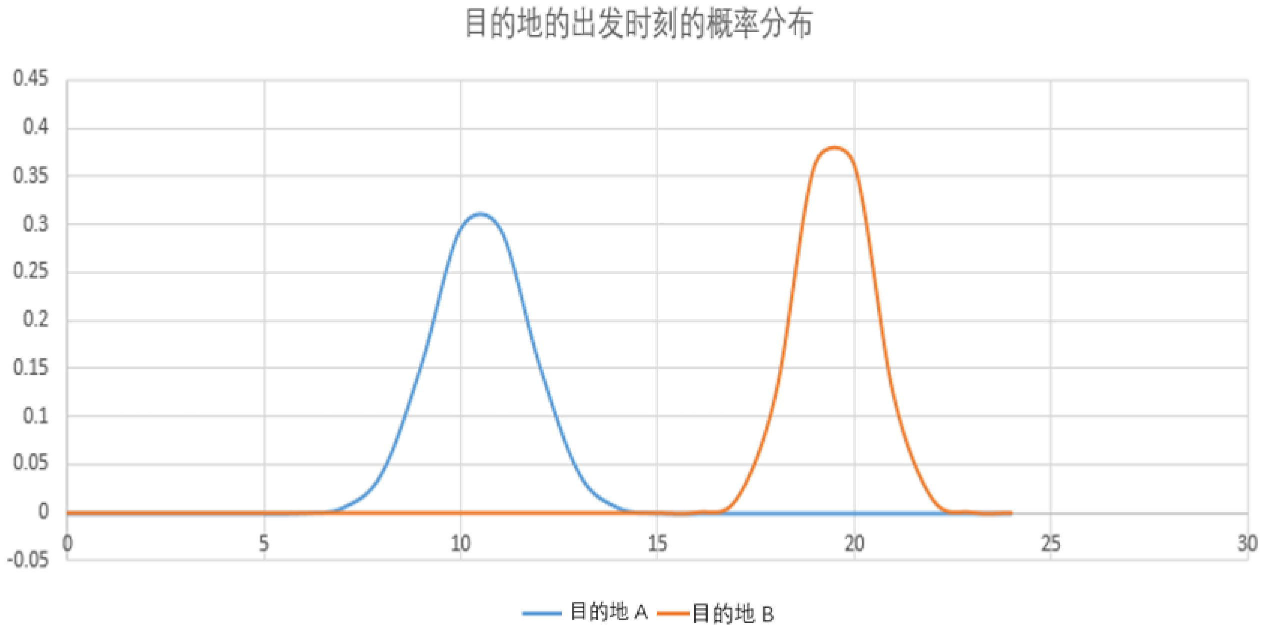
- 特征提取
  - 处理后的高维特征键可能不独立
- 离散化（0-1）
  - 非线性化
  - 提高性能

目录

Case1:出发时间区分不同目的地

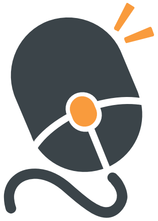
目的地 （D)	平均出发时刻	方差	出发时刻
目的地 A	10.5	1.24	8.7,9.7,9.9,9.9,9.9,10,10.1,10.1,10
目的地 B	19.3	0.96	18,18.2,18.9,19,19.3,20.5,21.1

Case1:出发时间区分不同目的地



目的地变量与出发时刻、出发地变量的互信息

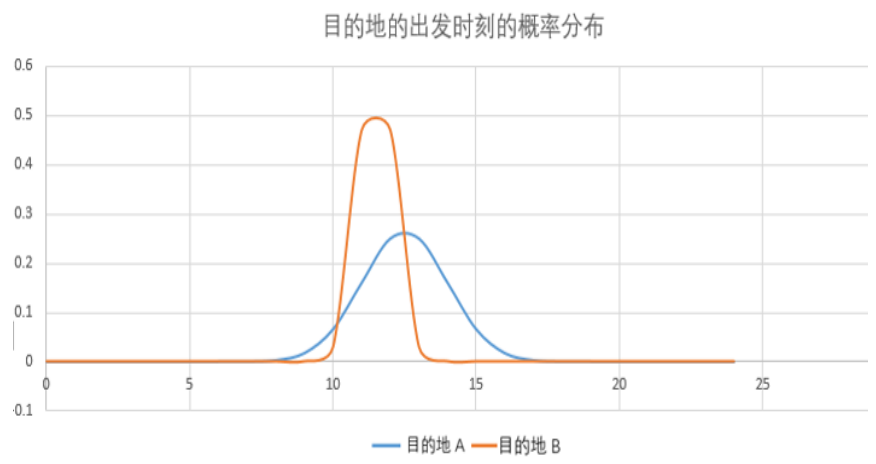
I(x, f)	I(x, t)	I(x, {f, t})
0.92	1.36	1.47

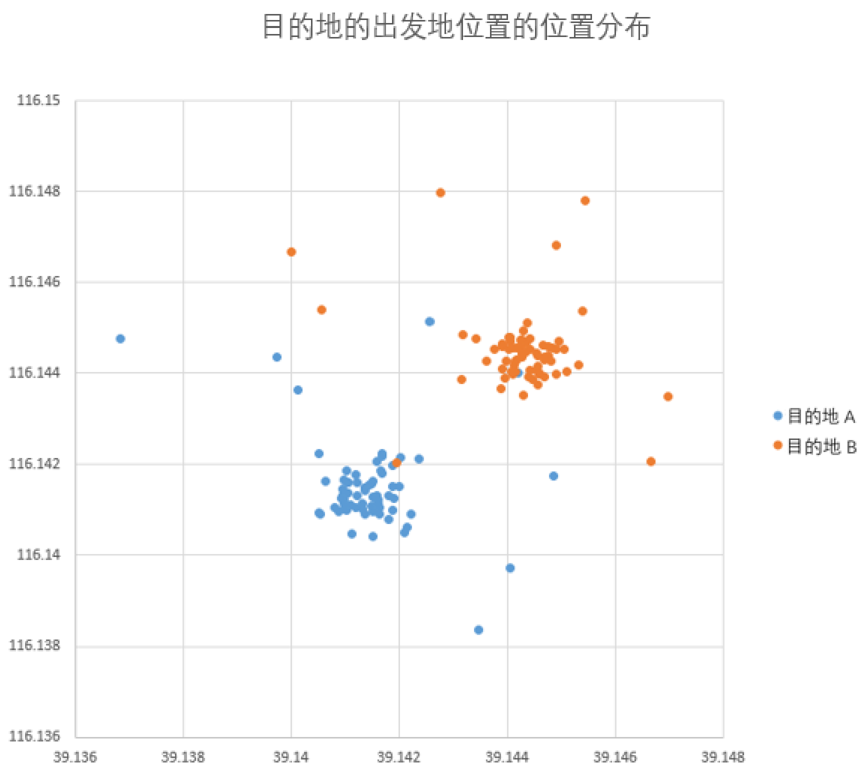


Case2:出发地区分不同目的地

目的地（D)	平均出发时刻	方差	出发时刻
目的地 A	12.4	1.5	13;11;13;10;13;13;13;14;14;13;12;9;13;12;10;9;
目的地 B	11.6	0.3	11;11;11;12;11;10;11;11;11;11;

Case2:出发地区分不同目的地



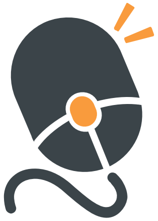


目的地变量与出发时刻、出发地变量的互信息

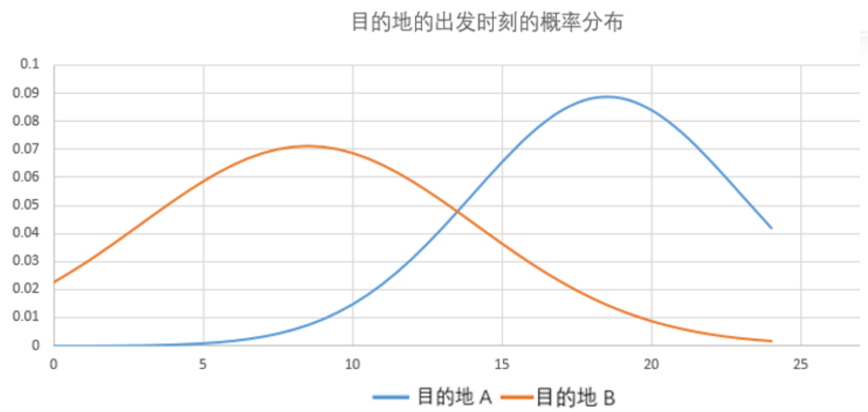
I(x, f)	I(x, t)	I(x, {f, t})
1.51	1.1	1.7

Case2:出发地区分不同目的地

出发地（Fn）	目的地 （D）	出发时间（T)
C	E	9:50:25
D	F	16:42:35
F	E	8:36:44
B	A	13:20:09
B	A	14:13:40
B	A	14:29:58
B	A	13:25:15
B	A	13:12:43
B	A	12:52:07
A	B	11:39:10
A	B	11:46:27
A	B	11:49:23
A	B	11:48:04
A	B	11:49:45
A	B	12:10:09
A	E	10:20:24
A	B	11:43:31
A	B	11:18:41
A	B	10:57:19
A	B	11:21:30
A	F	9:20:54
E	F	11:33:25

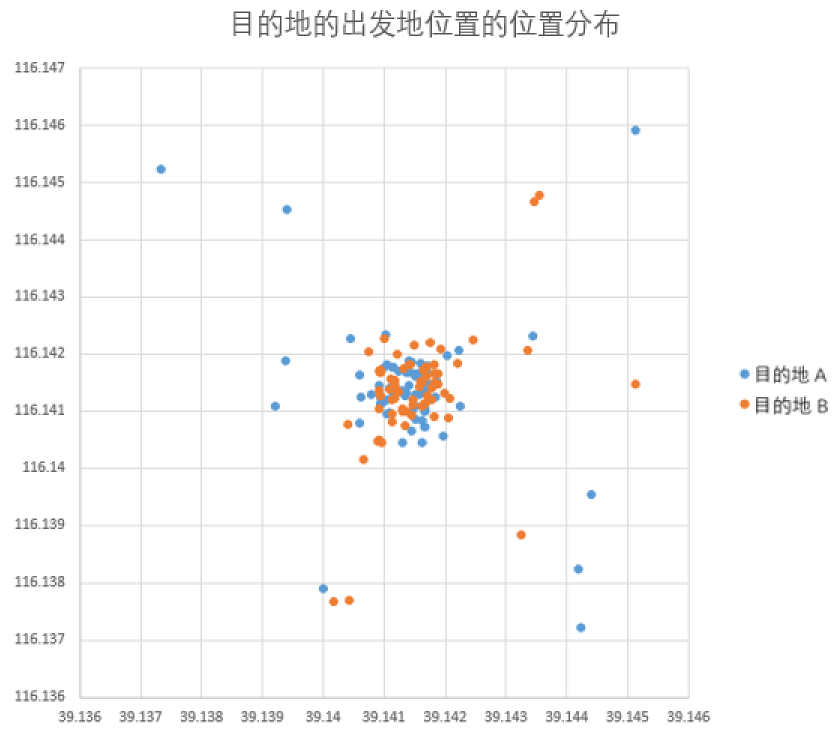


Case3:出发地和出发时间联合区分不同目的地



目的地变量与出发时刻、出发地变量的互信息

I(x, f)	I(x, t)	I(x, {f, t})
0.65	0.72	1.29



Case3:出发地和出发时间联合区分不同目的地

出发地	目的地	出发时间
C	B	12:34:26
D	B	7:45:19
E	B	18:18:00
F	B	1:45:08
F	B	8:02:29
A	J	13:37:37
A	I	11:38:02
A	H	13:39:36
A	G	9:40:18
A	B	18:01:00
A	B	17:11:51
A	B	18:00:49
A	B	18:03:48