My method for «dunnhumby's Shopper Challenge» problem solving
(http://www.kaggle.com/c/dunnhumbychallenge**)**

D'yakonov Alexander
(djakonov@mail.ru, http://alexanderdyakonov.narod.ru/)

Let's remind, that kernel density estimator of an unknown density of the sample $x_1,\dots,x_n \in \mathbf{R}$ is

$$\frac{1}{n}\sum_{i=1}^{n} K(x-x_i),$$

$$\text{where } K(x)=\frac{1}{10}\cdot\begin{cases}1, & |x|\le 10,\\ 0, & |x|>10.\end{cases}$$

The threshold 10 is selected from the contest rules (from the rules of evaluation). In more general case, with weights $w_1,\dots,w_n$:

$$f(x)=\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i K(x-x_i).$$

If $f_* = f(x_*) = \max[f(x)]$, then x_* is the mode, and f_* is the value of estimated probability density function f in the mode. In case of several modes we'll choose the median (middle point) of the modes.

We make the following assumptions for the date and dollar spend prediction for the customer:
1. The prediction depends only on statistics of the customer.
2. The customer will visit the store next week (i.e. it is necessary to select one of the following 7 days).

Let's delete from statistics of the customer weeks when he didn't visit the store; the probability of visit on the t-th day we'll estimate as

$$p_t = \frac{1}{48230}\sum_{r=1}^{52}(53-r)^2[0.875\cdot\delta(t-7\cdot r)+0.125\cdot\delta_1(t-7\cdot r)],$$

$\delta(s)=1 \Leftrightarrow$ the customer had visit on the s-th day (else $\delta(s)=0$),

$\delta_1(s)=1 \Leftrightarrow$ the customer had visit on the s-th day and it was the first visit that week,

we suppose that week begins from Friday (as Friday is the first day for prediction). Here, $48230=\sum_{r=1}^{52} r^2$, 0.875 is a parameter for tuning, $0.125=1-0.875$, 52 is the maximal number of weeks in statistics.

If we know probabilities of visit on the first day – p_1,
on the second day – p_2,

…

on the 7-th day – p_7,

then the probability that the first visit will be on the first day is $\tilde{p}_1 = p_1$,

on the second day – $\tilde{p}_2 = (1-p_1)p_2$,

…

on the 7-th day – $\tilde{p}_7 = \prod_{i=1}^{6}(1-p_i)p_7$.

It is natural for the date prediction to select

$$t : \tilde{p}_t = \max[\{\tilde{p}_t\}_{t=1}^{7}],$$

but we will multiply probabilities by «stability of the t-th type of the days». The stability of the t-th type is $m_t = f_*$, where $f(x)$ estimated from dollar spends on days of that type (on Fridays or on Saturdays or etc.) $v_1,\ldots,v_{n(t)}$ with weights $1,\ldots,n(t)$ (v_1 is the first purchase, $v_{n(t)}$ – the last purchase).

Day (for prediction) is selected by means of maximization of values

$$\{\tilde{p}_t \cdot (m_t + \varepsilon)\}_{t=1}^{7},$$

where $\varepsilon = 0.15$ is a parameter for tuning.

Now we predict dollar spend. We will write out the spends on days of that type (arranging from the recent spend to the first spend), no more than 40 (it is parameter for tuning):

$$v_1,\ldots,v_{n_1},$$

$n_1 = \min$(number of purchases on days of that type, 40).

Let's add the last purchases, no more than 6 + $0.4 \cdot n_1$ (parameters for tuning):

$$(v_1,\ldots,v_m) = (v_1,\ldots,v_{n_1},v'_1,\ldots,v'_6,v'_7,\ldots,v'_{n_2}),\ n_2 = \lfloor 0.4 \cdot n_1 \rfloor.$$

From this sample we estimate probability density function $f(x)$ using weights $\sqrt{m},\sqrt{m-1},\ldots,\sqrt{2},\sqrt{1}$. The mode x_* will be our prediction after «reducing to interval»:

$$x_* = \min(x_*, \text{round}(\max\{x\} - 10)),$$

$$x_* = \max(x_*, \text{round}(\min\{x\} + 10)),$$

where $\max\{x\}$ is the maximal spend (among all spends of the customer), $\min\{x\}$ is the minimal spend.