СТАТИСТИЧЕСКАЯ ТЕОРИЯ ПРИНЯТИЯ РЕШЕНИЙ

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург
27 января 2017 г.

Random facts:

- 27 января 1944 г. Ленинград был полностью освобождён от блокады;
- 27 января 1945 г. были освобождены узники Освенцима.

ФУНКЦИЯ ПОТЕРИ

- Сейчас мы попытаемся понять, что же на самом деле происходит в этих методах.
- Начнём с обычной регрессии – непрерывный вещественный вход $\mathbf{x} \in \mathbb{R}^p$, непрерывный вещественный выход $y \in \mathbb{R}$; у них есть некоторое совместное распределение $p(\mathbf{x}, y)$.
- Мы хотим найти функцию $f(\mathbf{x})$, которая лучше всего предсказывает y.

ФУНКЦИЯ ПОТЕРИ

- Введём *функцию потери* (loss function) $L(y, f(\mathbf{x}))$, которая наказывает за ошибки; естественно взять квадратичную функцию потери

$$L\left(y, f(\mathbf{x})\right) = (y - f(\mathbf{x}))^2.$$

- Тогда каждому f можно сопоставить *ожидаемую ошибку предсказания* (expected prediction error):

$$\mathrm{EPE}(f) = \mathrm{E}(y - f(\mathbf{x}))^2 = \int\int (y - f(\mathbf{x}))^2 p(\mathbf{x}, y) d\mathbf{x} dy.$$

- И теперь самая хорошая функция предсказания $\widehat{f}$ – это та, которая минимизирует $\mathrm{EPE}(f)$.

ФУНКЦИЯ ПОТЕРИ

- Это можно переписать как

$$\mathrm{EPE}(f) = \mathrm{E}_{\mathbf{x}}\mathrm{E}_{y|\mathbf{x}}\left[(y - f(\mathbf{x}))^2 \mid \mathbf{x}\right],$$

  и, значит, можно теперь минимизировать EPE поточечно:

$$\hat{f}(\mathbf{x}) = \arg\min_c \mathrm{E}_{y|\mathbf{x}'}\left[(y - c)^2 \mid \mathbf{x}' = \mathbf{x}\right],$$

  а это можно решить и получить

$$\hat{f}(\mathbf{x}) = \mathrm{E}_{y|\mathbf{x}'}(y \mid \mathbf{x}' = \mathbf{x}).$$

- Это решение называется *функцией регрессии* и является наилучшим предсказанием y в любой точке $\mathbf{x}$.

k-NN

- Теперь мы можем понять, что такое k-NN.
- Давайте оценим это ожидание:

$$f(\mathbf{x}) = \mathrm{E}_{y|\mathbf{x}'}(y \mid \mathbf{x}' = \mathbf{x}).$$

- Оценка ожидания – это среднее всех y с данным $\mathbf{x}$. Конечно, у нас таких нету, поэтому мы приближаем это среднее как

$$\hat{f}(\mathbf{x}) = \mathrm{Average}\left[y_i \mid \mathbf{x}_i \in N_k(\mathbf{x})\right].$$

- Это сразу два приближения: ожидание через среднее и среднее в точке через среднее в ближних точках.
- Иначе говоря, k-NN предполагает, что в окрестности $\mathbf{x}$ функция $y(\mathbf{x})$ не сильно меняется, а лучше всего – она кусочно-постоянна.

ЛИНЕЙНАЯ РЕГРЕССИЯ

- А линейная регрессия – это модельный подход, мы предполагаем, что функция регрессии линейна от своих аргументов:
$$f(\mathbf{x}) \approx \mathbf{x}^\top \mathbf{w}.$$
- Теперь мы не берём условие по $\mathbf{x}$, как в k-NN, а просто собираем много значений для разных $\mathbf{x}$ и обучаем модель.

BIAS-VARIANCE DECOMPOSITION

- Рассмотрим совместное распределение $p(y, \mathbf{x})$ и квадратичную функцию потерь $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$.
- Мы знаем, что тогда оптимальная оценка – это функция регрессии

$$\hat{f}(\mathbf{x}) = \mathrm{E}\left[y \mid \mathbf{x}\right] = \int y p(y \mid \mathbf{x}) dx.$$

BIAS-VARIANCE DECOMPOSITION

- Давайте подсчитаем ожидаемую ошибку и перепишем её в другой форме:

$$\mathrm{E}[L] = \mathrm{E}[(y - f(\mathbf{x}))^2] = \mathrm{E}[(y - \mathrm{E}\left[y \mid \mathbf{x}\right] + \mathrm{E}\left[y \mid \mathbf{x}\right] - f(\mathbf{x}))^2] =$$
$$= \int \left(f(\mathbf{x}) - \mathrm{E}\left[y \mid \mathbf{x}\right]\right)^2 p(\mathbf{x}) d\mathbf{x} + \int \left(\mathrm{E}\left[y \mid \mathbf{x}\right] - y\right)^2 p(\mathbf{x}, y) d\mathbf{x} dy,$$

потому что

$$\int \left(f(\mathbf{x}) - \mathrm{E}\left[y \mid \mathbf{x}\right]\right) \left(\mathrm{E}\left[y \mid \mathbf{x}\right] - y\right) p(\mathbf{x}, y) d\mathbf{x} dy = 0.$$

BIAS-VARIANCE DECOMPOSITION

- Эта форма записи – разложение на bias-variance и noise:

$$\mathrm{E}[L] = \int \left(f(\mathbf{x}) - \mathrm{E}\left[y \mid \mathbf{x}\right]\right)^2 p(\mathbf{x}) d\mathbf{x} + \int \left(\mathrm{E}\left[y \mid \mathbf{x}\right] - y\right)^2 p(\mathbf{x}, y) d\mathbf{x} dy,$$

- Отсюда, кстати, тоже сразу видно, что от $f(\mathbf{x})$ зависит только первый член, и он минимизируется, когда

$$f(\mathbf{x}) = \hat{f}(\mathbf{x}) = \mathrm{E}\left[y \mid \mathbf{x}\right].$$

- А noise, $\int \left(\mathrm{E}\left[y \mid \mathbf{x}\right] - y\right)^2 p(\mathbf{x}, y) d\mathbf{x} dy$, – это просто свойство данных, дисперсия шума.

BIAS-VARIANCE DECOMPOSITION

- Если бы у нас был всемогущий компьютер и неограниченный датасет, мы бы, конечно, на этом и закончили, посчитали бы $\hat{f}(\mathbf{x}) = \mathrm{E}\left[y \mid \mathbf{x}\right]$, и всё.
- Однако жизнь – борьба, и у нас есть только ограниченный датасет из N точек. Предположим, что этот датасет берётся по распределению $p(\mathbf{x}, y)$ – т.е. фактически рассмотрим много-много экспериментов такого вида:
  - взяли датасет D из N точек по распределению $p(\mathbf{x}, y)$;
  - подсчитали нашу чудо-регрессию;
  - получили новую функцию предсказания $f(\mathbf{x}; D)$.
- Разные датасеты будут приводить к разным функциям предсказания…

- ...а потому давайте усредним теперь по датасетам.
- Наш первый член в ожидаемой ошибке выглядел как $\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2$, а теперь будет $\left(f(\mathbf{x}; D) - \hat{f}(\mathbf{x})\right)^2$, и его можно усреднить по D, применив такой же трюк:

$$\left(f(\mathbf{x}; D) - \hat{f}(\mathbf{x})\right)^2$$
$$= \left(f(\mathbf{x}; D) - \mathrm{E}_D\left[f(\mathbf{x}; D)\right] + \mathrm{E}_D\left[f(\mathbf{x}; D)\right] - \hat{f}(\mathbf{x})\right)^2$$
$$= \left(f(\mathbf{x}; D) - \mathrm{E}_D\left[f(\mathbf{x}; D)\right]\right)^2 + \left(\mathrm{E}_D\left[f(\mathbf{x}; D)\right] - \hat{f}(\mathbf{x})\right)^2 + 2(\ldots)(\ldots),$$

и в ожидании получится...

BIAS-VARIANCE DECOMPOSITION

- ...и в ожидании получится

$$\mathrm{E}_D\left[\left(f(\mathbf{x};D)-\hat{f}(\mathbf{x})\right)^2\right]=$$
$$=\mathrm{E}_D\left[\left(f(\mathbf{x};D)-\mathrm{E}_D\left[f(\mathbf{x};D)\right]\right)^2\right]+\left(\mathrm{E}_D\left[f(\mathbf{x};D)\right]-\hat{f}(\mathbf{x})\right)^2.$$

- Разложили на дисперсию $\mathrm{E}_D\left[\left(f(\mathbf{x};D)-\mathrm{E}_D\left[f(\mathbf{x};D)\right]\right)^2\right]$ и квадрат систематической ошибки $\left(\mathrm{E}_D\left[f(\mathbf{x};D)\right]-\hat{f}(\mathbf{x})\right)^2$; это и есть bias-variance decomposition.

BIAS-VARIANCE-NOISE

$$\text{Expected loss} = (\text{bias})^2 + \text{variance} + \text{noise},$$

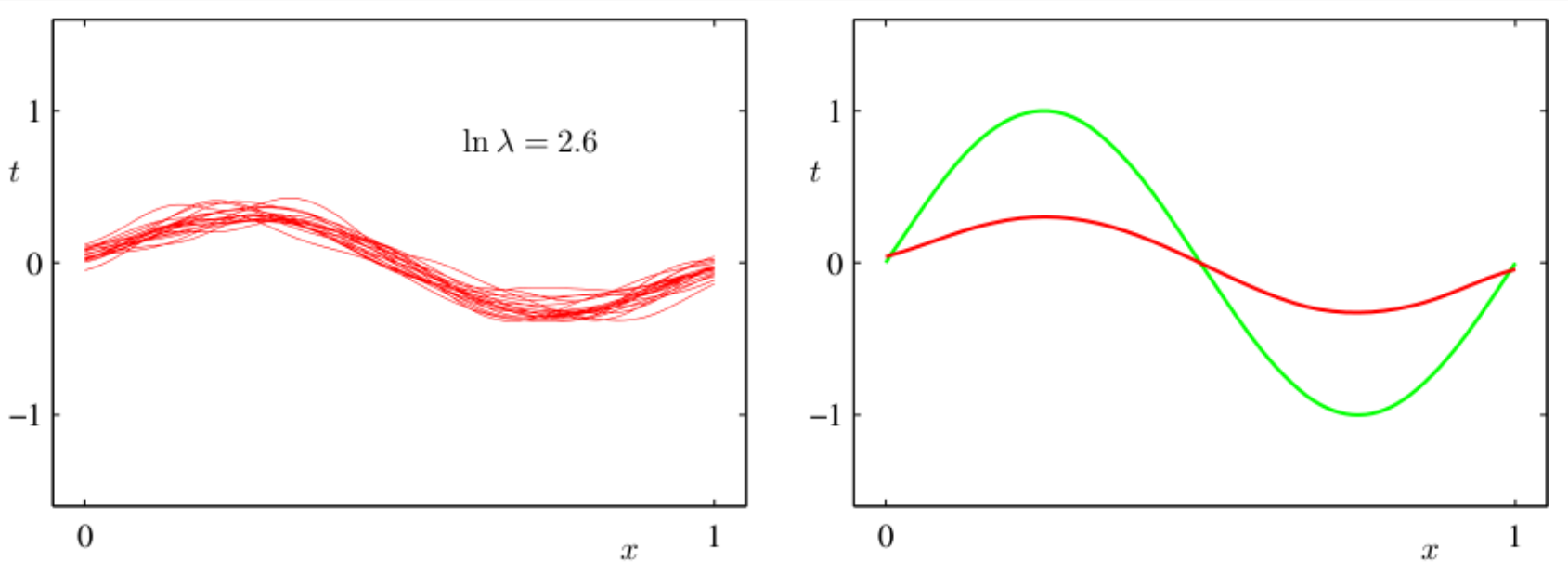
где

$$(\text{bias})^2 = \left(\mathrm{E}_D\left[f(\mathbf{x}; D)\right] - \hat{f}(\mathbf{x})\right)^2,$$
$$\text{variance} = \mathrm{E}_D\left[\left(f(\mathbf{x}; D) - \mathrm{E}_D\left[f(\mathbf{x}; D)\right]\right)^2\right],$$
$$\text{noise} = \int \left(\mathrm{E}\left[y \mid \mathbf{x}\right] - y\right)^2 p(\mathbf{x}, y) d\mathbf{x} dy.$$

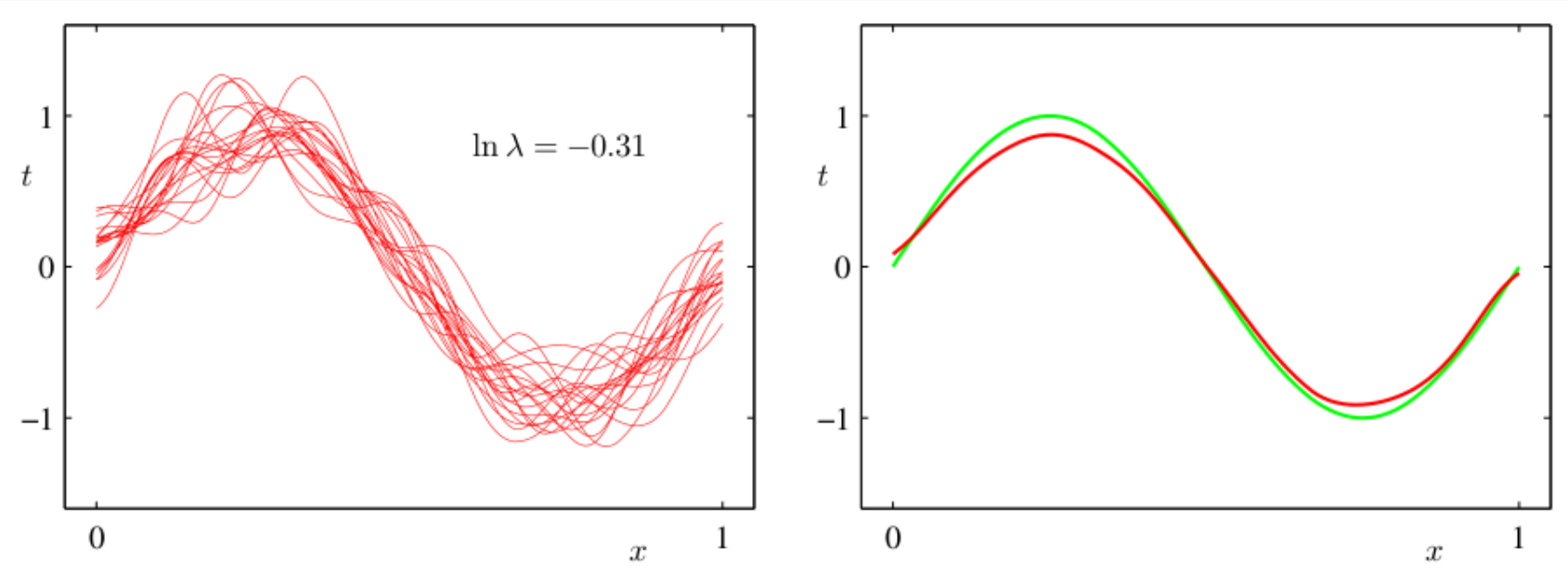
ПРИМЕР

- Теперь давайте посмотрим на пример: опять та же синусоида, опять приближаем её линейной регрессией с полиномиальными признаками (максимальным их числом).
- И мы регуляризуем эту регрессию с параметром α.
- Будем набрасывать много датасетов и смотреть, что меняется при этом.

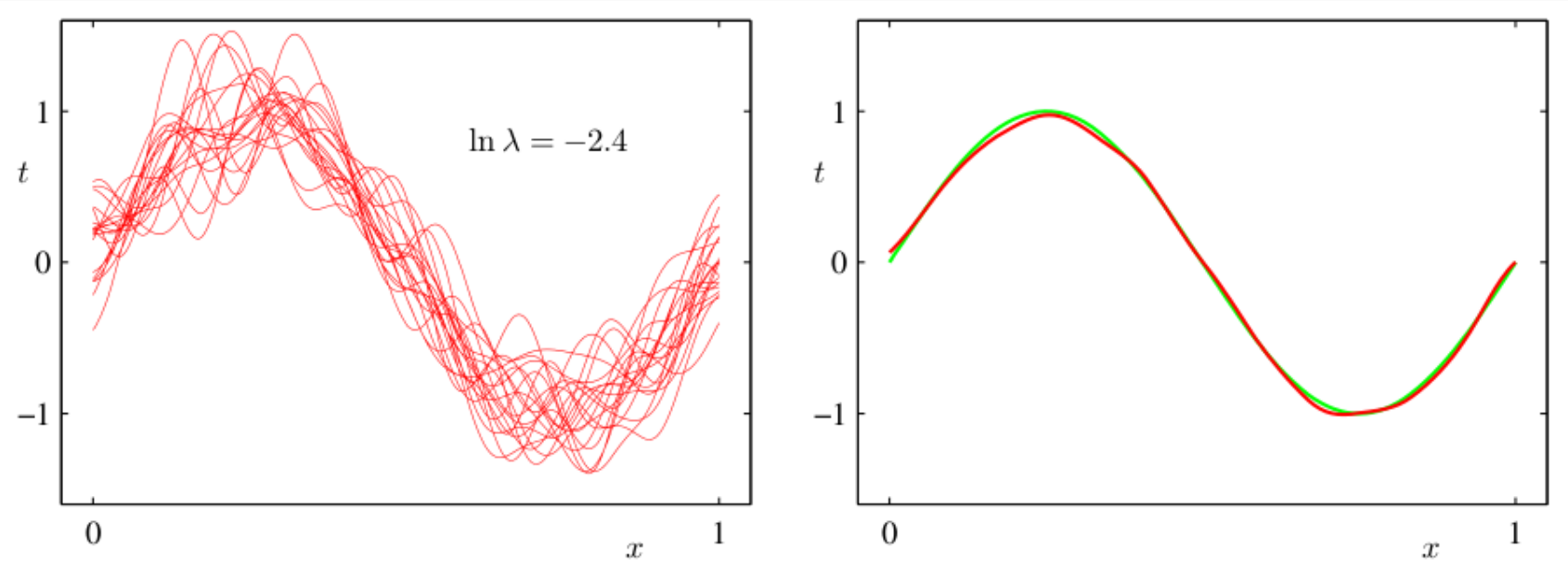
РЕГУЛЯРИЗАТОР И BIAS-VARIANCE



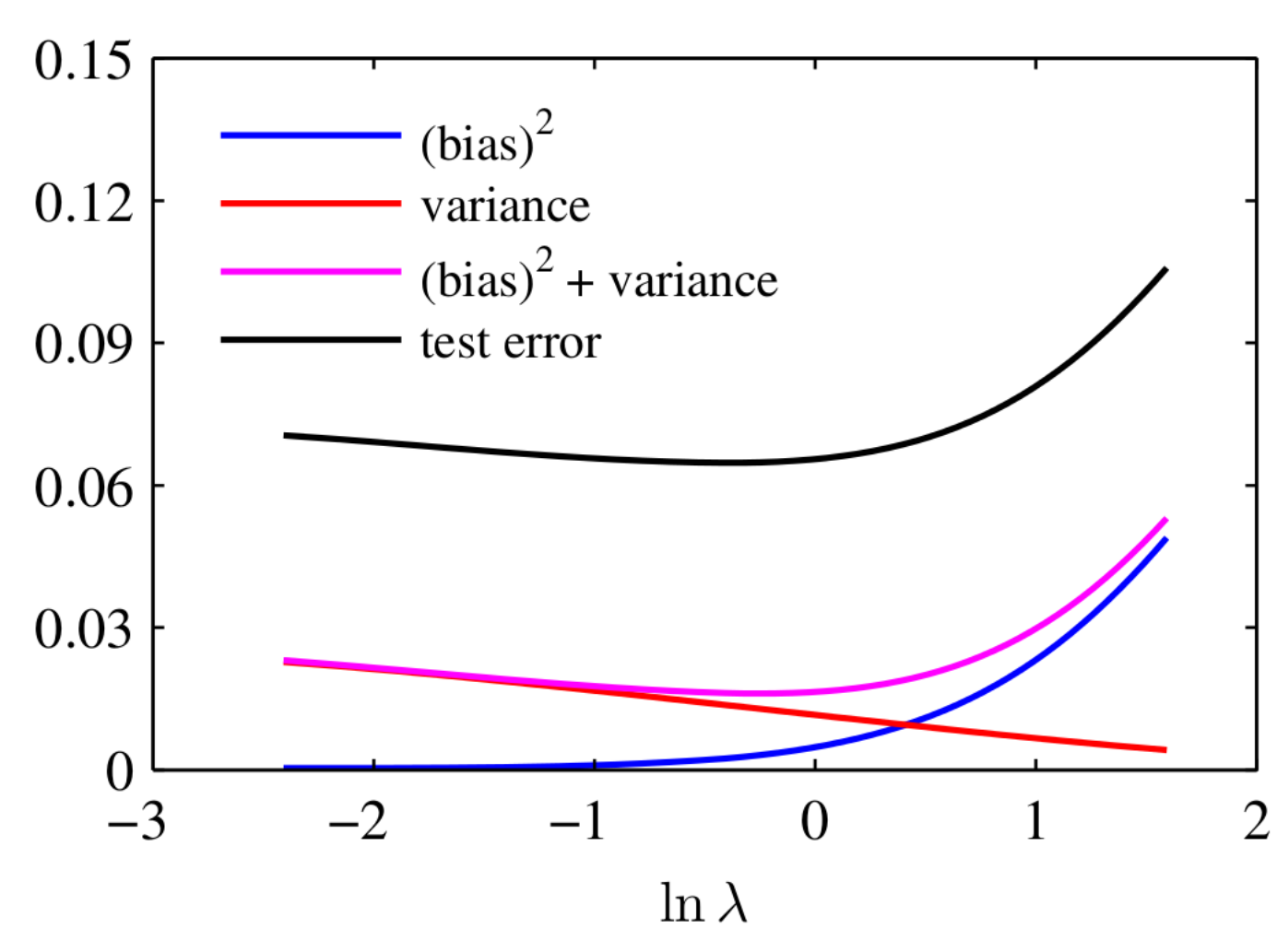
РЕГУЛЯРИЗАТОР И BIAS-VARIANCE



РЕГУЛЯРИЗАТОР И BIAS-VARIANCE



РЕГУЛЯРИЗАТОР И BIAS-VARIANCE



СПАСИБО!

Спасибо за внимание!