

Линейная регрессия: метод наименьших квадратов

Сергей Николенко

Академический Университет, 2012

Outline

- 1 Наименьшие квадраты и ближайшие соседи
 - Метод наименьших квадратов
 - Метод ближайших соседей
- 2 Статистическая теория принятия решений
 - Регрессия
 - Классификация
- 3 О регрессии по-байесовски
 - Нормальное распределение
 - Байесовская регрессия

В предыдущей серии...

- Теорема Байеса:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- Две основные задачи байесовского вывода:

- 1 найти апостериорное распределение на гипотезах/параметрах:

$$p(\theta | D) \propto p(D|\theta)p(\theta)$$

(и/или найти гипотезу максимального правдоподобия $\arg \max_{\theta} p(\theta | D)$);

- 2 найти апостериорное распределение исходов дальнейших экспериментов:

$$p(x | D) \propto \int_{\theta \in \Theta} p(x | \theta)p(D|\theta)p(\theta)d\theta.$$

Метод наименьших квадратов

- Линейная модель: рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^p x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \dots, x_p).$$

- Таким образом, по вектору входов $\mathbf{x}^\top = (x_1, \dots, x_p)$ мы будем предсказывать выход y как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^p x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

Метод наименьших квадратов

- Как найти оптимальные параметры $\hat{\mathbf{w}}$ по тренировочным данным вида $(\mathbf{x}_i, y_i)_{i=1}^N$?
- Метод наименьших квадратов: будем минимизировать

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^N (y_i - \mathbf{x}_i^T \mathbf{w})^2.$$

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

$$\text{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

где \mathbf{X} – матрица $N \times p$, продифференцировать по \mathbf{w} , получится

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

если матрица $\mathbf{X}^\top \mathbf{X}$ невырожденная.

- Замечание: $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы \mathbf{X} ; это обобщение понятия обратной матрицы на неквадратные матрицы.
- Много ли нужно точек, чтобы обучить такую модель?

Метод наименьших квадратов

- Пример: задача классификации. Два класса; мы кодируем один ответ как $y = 1$, другой ответ как $y = 0$ и рисуем прямую $\mathbf{x}^\top \hat{\mathbf{w}} = \frac{1}{2}$.
- Мы видим, что данные разделяются не то чтобы совсем замечательно.
- Когда линейная модель работает хорошо, когда плохо?
- Предположим, что это была смесь нескольких нормальных распределений – что тогда?

Метод ближайших соседей

- Линейная модель – очень сильные предположения, много точек не нужно.
- Совсем другой подход – давайте вообще никаких предположений не делать (это не совсем так, конечно :)), а будем отталкиваться от данных.
- Давайте не будем строить вообще никакой модели, а будем классифицировать новые примеры как

$$\hat{y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i,$$

где $N_k(\mathbf{x})$ – множество k ближайших соседей точки \mathbf{x} среди имеющихся данных $(\mathbf{x}_i, y_i)_{i=1}^N$.

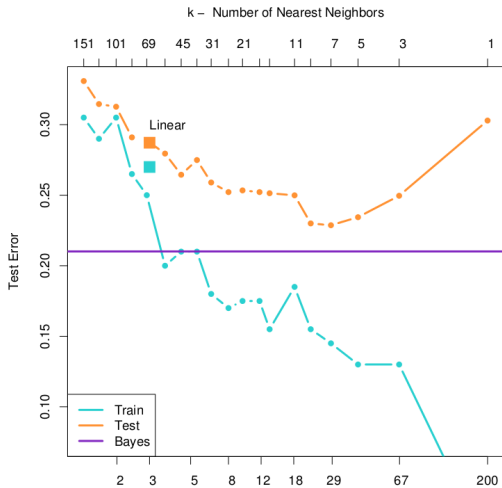
Метод ближайших соседей

- Снова смотрим на примеры – теперь появился параметр k , от которого многое зависит.
- Для разумно большого k у нас в нашем примере стало меньше ошибок.
- Но это не предел – для $k = 1$ на тестовых данных вообще никаких ошибок нету!
- Что это значит? В чём недостаток метода ближайших соседей при $k = 1$?
- Сколько параметров у метода k -NN?
- Как выбрать k ? Можно ли просто подсчитать среднеквадратическую ошибку и минимизировать её?

Метод ближайших соседей

- На самом деле данные были порождены так:
 - сначала по распределению $\mathcal{N}((1, 0)^T, \mathbf{I})$ породили 10 синих средних;
 - потом по распределению $\mathcal{N}((0, 1)^T, \mathbf{I})$ породили 10 красных средних;
 - потом для каждого из классов сгенерировали по 100 точек так: выбрать одно из 10 средних m_k равномерно (с вероятностью $\frac{1}{10}$), потом породили точку $\mathcal{N}(m_k, \frac{1}{5}\mathbf{I})$.
- Получилось, что мы разделяем две смеси гауссианов.

Качество метода K -NN



Outline

- 1 Наименьшие квадраты и ближайшие соседи
 - Метод наименьших квадратов
 - Метод ближайших соседей
- 2 Статистическая теория принятия решений
 - Регрессия
 - Классификация
- 3 О регрессии по-байесовски
 - Нормальное распределение
 - Байесовская регрессия

Функция потерь

- Сейчас мы попытаемся понять, что же на самом деле происходит в этих методах.
- Начнём с настоящей 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 потерь

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

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

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

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

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

Функция потерь

- Это можно переписать как

$$EPE(f) = E_{\mathbf{x}} E_{y|\mathbf{x}} [(y - f(\mathbf{x}))^2 | \mathbf{x}],$$

и, значит, можно теперь минимизировать EPE поточечно:

$$\hat{f}(\mathbf{x}) = \arg \min_c E_{y|\mathbf{x}'} [(y - c)^2 | \mathbf{x}' = \mathbf{x}],$$

а это можно решить и получить

$$\hat{f}(\mathbf{x}) = E_{y|\mathbf{x}'}(y | \mathbf{x}' = \mathbf{x}).$$

- 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:

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

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

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

- 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:

$$f(\mathbf{x}) \approx \mathbf{x}^T \mathbf{w}.$$

- Теперь мы не берём условие по x , как в k -NN, а просто собираем много значений для разных \mathbf{x} и обучаем модель.

Классификация

- То же самое можно и с задачей классификации сделать. Пусть у нас переменная g с K возможными значениями g_1, \dots, g_k предсказывается.
- Введём функцию потери, равную 1 за каждый неверный ответ. Получим

$$\text{EFE} = \mathbb{E}[L(g, \hat{g}(\mathbf{x}))].$$

- Перепишем как раньше:

$$\text{EFE} = \mathbb{E}_{\mathbf{x}} \sum_{k=1}^K L(g_k, \hat{g}(\mathbf{x})) p(g_k | \mathbf{x}).$$

- Опять достаточно оптимизировать поточечно:

$$\hat{g}(\mathbf{x}) = \arg \min_g \sum_{k=1}^K L(g_k, \hat{g}(\mathbf{x})) p(g_k | \mathbf{x}).$$

Классификация

- Опять достаточно оптимизировать поточечно:

$$\hat{g}(\mathbf{x}) = \arg \min_g \sum_{k=1}^K L(g_k, \hat{g}(\mathbf{x})) p(g_k | \mathbf{x}).$$

- Для 0-1 функции потери это упрощается до

$$\hat{g}(\mathbf{x}) = \arg \min_g [1 - p(g | \mathbf{x})], \text{ т.е.}$$

$$\hat{g}(\mathbf{x}) = g_k, \text{ если } p(g_k | \mathbf{x}) = \max_g p(g | \mathbf{x}).$$

- Это называется *оптимальным байесовским классификатором*; если модель известна, то его обычно можно построить.

Классификация

- Байесовский классификатор: $\hat{g}(\mathbf{x}) = g_k$ для $p(g_k | \mathbf{x}) = \max_g p(g | \mathbf{x})$.
- Опять k -NN строит приближение к этой формуле – выбирает большинством голосов в окрестности точки.
- Что делает линейный классификатор, мы уже обсуждали – кодируем g через 0-1 переменную y , приближаем y линейной функцией, предсказываем.
- Правда, странно получается – наше приближение может быть отрицательным или большим 1, например.

Outline

- 1 Наименьшие квадраты и ближайшие соседи
 - Метод наименьших квадратов
 - Метод ближайших соседей
- 2 Статистическая теория принятия решений
 - Регрессия
 - Классификация
- 3 О регрессии по-байесовски
 - Нормальное распределение
 - Байесовская регрессия

Поиск скрытых параметров

- Сначала – небольшое лирическое отступление о нормальном распределении. Кстати, почему все всё время предполагают нормальное распределение?
- Очень многие задачи машинного обучения можно представить как *поиск скрытых параметров*.
- Есть некоторое предположение о структуре задачи, т.е. о виде распределений, которыми набрасываются тестовые данные.
- Требуется найти наиболее правдоподобные неизвестные параметры этих распределений.

Гауссиан

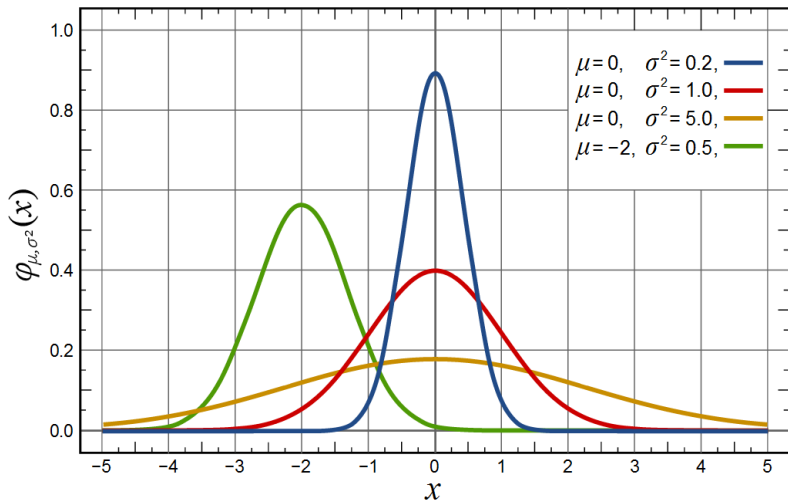
- Давайте решим эту задачу для нормального (гауссовского) распределения. У него два параметра:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- Функция правдоподобия данных x_1, \dots, x_n :

$$p(x_1, \dots, x_n | \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}}.$$

Нормальное распределение



Гауссиан: достаточные статистики

- Заметим, что функция эта зависит от двух параметров, а не от n :

$$p(x_1, \dots, x_n | \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{S+n(\bar{x}-\mu)^2}{2\sigma^2}},$$

где

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}, \quad S = \sum_{i=1}^n (\bar{x} - x_n)^2.$$

- Параметры \bar{x} и S называются *достаточными статистиками* (sufficient statistics).

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \dots, x_n | \mu, \sigma) = -n \ln(\sigma \sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \dots, x_n | \mu, \sigma) = -n \ln(\sigma \sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?
- Взять частные производные и приравнять нулю.

Гауссиан: ГМП

- По μ :

$$\frac{\partial \ln p}{\partial \mu} = -\frac{n}{\sigma^2}(\mu - \bar{x}).$$

- То есть в гипотезе максимального правдоподобия $\mu_{ML} = \bar{x}$, независимо от S .
- Теперь нужно найти σ из гипотезы максимального правдоподобия.
- Для этого мы продифференцируем по $\ln \sigma$ — полезный приём на будущее. Кстати, $\frac{dx^n}{d(\ln x)} = nx^n$.

Гауссиан: ГМП



$$\frac{\partial \ln p}{\partial \ln \sigma} = -n + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2}.$$

- Следовательно, в гипотезе максимального правдоподобия

$$\sigma_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{ML})^2.$$

- Интересно, что это *смещённая* оценка, т.е. ожидание σ_{ML}^2 по распределению $\mathcal{N}(\mu, \sigma^2)$ не равно σ , а несмещённая оценка получается как

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2.$$

Несколько гауссианов

- Теперь то же самое для нескольких гауссианов сразу.
- Даны несколько точек x_1, \dots, x_n , но они принадлежат смеси гауссианов с разными μ_k и σ_k .
- Обозначим коэффициенты смеси через w_k (вероятность того, что точка порождена гауссианом со средним μ_k).
- Тогда распределение будет

$$p(x|\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_k w_k \mathcal{N}(x | \mu_k, \sigma_k^2).$$

Несколько гауссианов

- Мы будем ещё говорить о том, как находить параметры максимального правдоподобия в смесях распределений вроде этой;

$$p(x_1, \dots, x_n | \mathbf{w}, \boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^n p(x_i | \mathbf{w}, \boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^n \sum_k w_k \mathcal{N}(x_i | \mu_k, \sigma_k^2).$$

- А пока просто – что мы сделаем, если найдём эти параметры? Как называется эта задача?

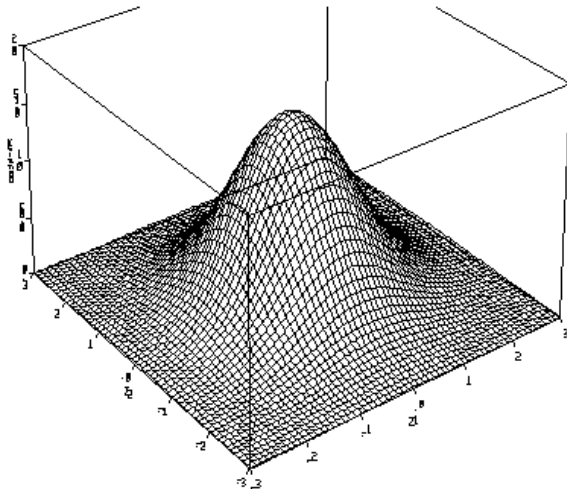
Несколько гауссианов

- Мы будем ещё говорить о том, как находить параметры максимального правдоподобия в смесях распределений вроде этой;

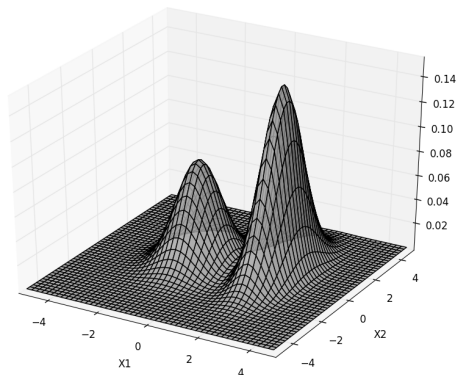
$$p(x_1, \dots, x_n | \mathbf{w}, \boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^n p(x_i | \mathbf{w}, \boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^n \sum_k w_k \mathcal{N}(x_i | \mu_k, \sigma_k^2).$$

- А пока просто – что мы сделаем, если найдём эти параметры? Как называется эта задача?
- Мы решим задачу *кластеризации*: разделим данные (по наиболее правдоподобной компоненте смеси) на хорошо определённые (в данном случае гауссовские) кластеры.

Двумерный гауссиан



Смесь двумерных гауссианов



Байесовская регрессия

- Теперь давайте поговорим о линейной регрессии по-байесовски.
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е. переменная t , которую мы наблюдаем, получается как

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Иными словами,

$$p(t | \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t | y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- Здесь пока y – любая функция.

Байесовская регрессия

- Чтобы не повторять совсем уж то же самое, мы рассмотрим не в точности линейную регрессию, а её естественное обобщение – линейную модель с базисными функциями:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\Phi}(\mathbf{x})$$

(M параметров, $M - 1$ базисная функция, $\phi_0(\mathbf{x}) = 1$).

Байесовская регрессия

- Базисные функции ϕ_j – это, например:
 - результат feature extraction;
 - расширение линейной модели на нелинейные зависимости (например, $\phi_j(x) = x^j$);
 - локальные функции, которые существенно не равны нулю только в небольшой области (например, гауссовские базисные функции $\phi_j(\mathbf{x}) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
 - ...

Байесовская регрессия

- Рассмотрим набор данных $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ со значениями $\mathbf{t} = \{t_1, \dots, t_N\}$.
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\Phi}(\mathbf{x}_n), \sigma^2).$$

- Прологарифмируем (опустим \mathbf{X} , т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} | \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N \left(t_n - \mathbf{w}^\top \boldsymbol{\Phi}(\mathbf{x}_n)\right)^2.$$

Байесовская регрессия

- Прологарифмируем (опустим \mathbf{X} , т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} | \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N \left(t_n - \mathbf{w}^\top \boldsymbol{\Phi}(\mathbf{x}_n) \right)^2.$$

- И вот мы получили, что для максимизации правдоподобия по \mathbf{w} нам нужно как раз минимизировать среднеквадратичную ошибку!

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} | \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^N \left(t_n - \mathbf{w}^\top \boldsymbol{\Phi}(\mathbf{x}_n) \right) \boldsymbol{\Phi}(\mathbf{x}_n).$$

Байесовская регрессия

- Решая систему уравнений $\nabla \ln p(\mathbf{t} | \mathbf{w}, \sigma^2) = 0$, получаем то же самое, что и раньше:

$$\mathbf{w}_{ML} = \left(\Phi^T \Phi \right)^{-1} \Phi^T \mathbf{t}.$$

- Здесь $\Phi = (\phi_j(\mathbf{x}_i))_{i,j}$.

Байесовская регрессия

- Теперь можно и относительно σ^2 максимизировать правдоподобие; получим

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N \left(t_n - \mathbf{w}_{ML}^\top \boldsymbol{\Phi}(\mathbf{x}_n) \right)^2,$$

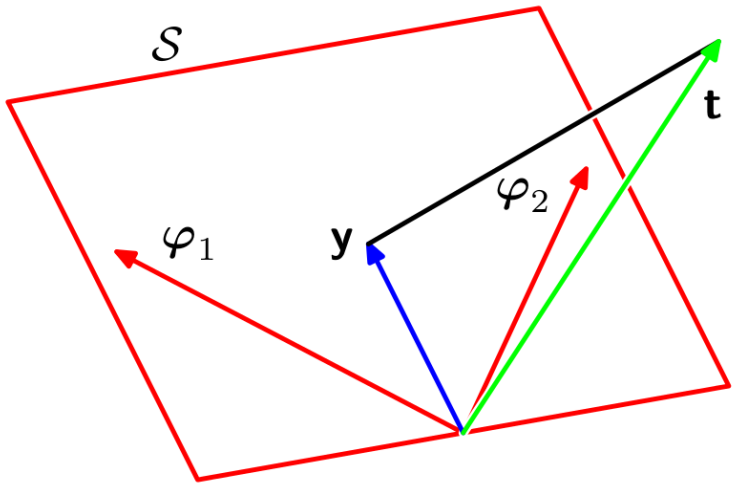
т.е. как раз выборочная дисперсия имеющихся данных вокруг предсказанного значения.

Геометрическая интерпретация

- Метод наименьших квадратов можно интерпретировать геометрически.
- Рассмотрим N -мерное пространство, в котором оси соответствуют t_n , и \mathbf{t} – вектор в нём.
- Если базисных функций ϕ_j меньше, чем точек \mathbf{x}_n , то векторы $\phi_j(\mathbf{x}_n)$ будут выделять какое-то подпространство в этом N -мерном пространстве.

Упражнение. Докажите, что метод наименьших квадратов (максимальное правдоподобие для нормально распределённого шума) эквивалентен проецированию вектора \mathbf{t} на подпространство $\langle \phi_0(\mathbf{x}), \dots, \phi_{M-1}(\mathbf{x}) \rangle$.

Геометрическая интерпретация



Thank you!

Спасибо за внимание!