香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 4

Generalized Linear Models, Kernel Methods

Junxian He
Feb 24, 2026

Announcement

HW1 is out, due on March 3rd, please start early

Exponential Family

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

$$P(y;\eta) = b(y) \exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

$\log \exp[a(\eta)] = a(\eta)$

$T(y)$ is called the **sufficient statistic**. Σ

$b(y)$ is called the **base measure** – does *not* depend on η.

$a(\eta)$ is called the **log partition function** – does *not* depend on y.

$\sum_y P(y) = 1$

partition

$$P(y) = \frac{b(y)\exp\{\eta^T T(y)\}}{\exp[a(\eta)]}$$

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

$T(y) = y$

$T(y)$ is called the **sufficient statistic**. holds all information the data provides with regard to the unknown parameter values

$b(y)$ is called the **base measure** – does *not* depend on η.

$a(\eta)$ is called the **log partition function** – does *not* depend on y.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

$T(y)$ is called the **sufficient statistic**. holds all information the data provides with regard to the unknown parameter values

$b(y)$ is called the **base measure** – does *not* depend on η.

$a(\eta)$ is called the **log partition function** – does *not* depend on y.

$$1 = \sum_y P(y;\eta) = e^{-a(\eta)} \sum_y b(y)\exp\left\{\eta^T T(y)\right\}$$

$$\implies a(\eta) = \log \sum_y b(y)\exp\left\{\eta^T T(y)\right\}$$

Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

discrete binary 0, 1

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

$y = \{0, 1\}$

$\phi = P(y=1)$

Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

$$p(y; \phi) = \phi^y (1-\phi)^{1-y}$$

How do we put it in the required form?

$$P(y; \eta) = b(y) \exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

$$p(y;\phi) = \phi^y(1-\phi)^{1-y}$$

How do we put it in the required form?

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

$b(y) = 1$

$T(y) = y$

$\eta = \log\frac{\phi}{1-\phi}$

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

Example: Bernoulli

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}$$

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

Example: Bernoulli

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}$$

$$\begin{aligned}
p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\
&= \exp(y\log\phi + (1-y)\log(1-\phi)) \\
&= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right)
\end{aligned}$$

So then:

$$\eta = \log\frac{\phi}{1-\phi},\ T(y) = y,\ a(\eta) = -\log(1-\phi).$$

$$b(y) = 1$$

$a(\eta) = -\log(1-\phi)$

$\eta = \log\frac{\phi}{1-\phi}$

Example: Bernoulli

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}$$

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

So then:

$$\eta = \log\frac{\phi}{1-\phi},\ T(y) = y,\ a(\eta) = -\log(1-\phi).$$

$$b(y) = 1$$

We need to show $a(\eta)$ is a function of $\log\frac{\phi}{1-\phi}$

Example: Bernoulli

Example: Bernoulli

We first observe that:

$$\eta = \log \frac{\phi}{1-\phi} \implies e^{\eta}(1-\phi) = \phi$$

$$e^{\eta} = (e^{\eta}+1)\phi \implies \phi = \frac{1}{1+e^{-\eta}}$$

Example: Bernoulli

We first observe that:

$$\phi = f(\eta) = e^{\eta}$$

$$\eta = \log \frac{\phi}{1-\phi} \implies e^{\eta}(1-\phi) = \phi$$

$$e^{\eta} = (e^{\eta}+1)\phi \implies \phi = \frac{1}{1+e^{-\eta}}$$

$$\phi = \frac{1}{1+e^{-\eta}}$$

Now, we plug into $\log(1-\phi)$ and we verify:

$$a(\eta) = \log(1-\phi) = \log \frac{e^{-\eta}}{1+e^{-\eta}} = -\log(1+e^{\eta}).$$

Example: Bernoulli

We first observe that:

$$\eta = \log \frac{\phi}{1-\phi} \implies e^{\eta}(1-\phi) = \phi$$
$$e^{\eta} = (e^{\eta}+1)\phi \implies \phi = \frac{1}{1+e^{-\eta}}$$

Now, we plug into $\log(1-\phi)$ and we verify:

$$a(\eta) = \log(1-\phi) = \log \frac{e^{-\eta}}{1+e^{-\eta}} = -\log(1+e^{\eta}).$$

We have verified Bernoulli distribution is in the exponential family

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}.$$

Can we put it in the exponential family form?

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}.$$

Can we put it in the exponential family form?

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

Multiply out the square and group terms:

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-y^2/2\right\} \exp\left\{\mu y - \frac{1}{2}\mu^2\right\}.$$

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}.$$

Can we put it in the exponential family form?

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

Multiply out the square and group terms:

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-y^2/2\right\} \exp\left\{\mu y - \frac{1}{2}\mu^2\right\}.$$

$$\eta = \mu, \ T(y) = y, \ a(\eta) = \frac{1}{2}\eta^2.$$

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}.$$

Can we put it in the exponential family form?

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

Multiply out the square and group terms:

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-y^2/2\right\} \exp\left\{\mu y - \frac{1}{2}\mu^2\right\}.$$

$$\eta = \mu,\ T(y) = y,\ a(\eta) = \frac{1}{2}\eta^2.$$

In all the exponential family distribution we work with in the course, T(y) = y

An Observation

An Observation

Notice that for a Gaussian with mean μ we had

$$\eta = \mu,\ T(y) = y,\ a(\eta) = \frac{1}{2}\eta^2.$$

An Observation

Notice that for a Gaussian with mean μ we had

$$\eta = \mu,\ T(y) = y,\ a(\eta) = \frac{1}{2}\eta^2.$$

natural parameter

$a(\eta) = \frac{1}{2}\eta^2$

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial_\eta^2 a(\eta) = 1 = \sigma^2 = \mathrm{var}(y)$$

$\frac{\partial a(\eta)}{\partial \eta} = \mu = E[y]$

$\partial_\eta^2 a(\eta) = 1 = \mathrm{Var}(y)$

An Observation

Notice that for a Gaussian with mean μ we had

$$\eta = \mu,\ T(y) = y,\ a(\eta) = \frac{1}{2}\eta^2.$$

$E[y] = \sum_y P(y)\, y$

$Var(y) = E[(y - \mu)^2]$

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial_\eta^2 a(\eta) = 1 = \sigma^2 = \mathrm{var}(y)$$

Is this true for general?

Log Partition Function

Yes! Recall that

$$a(\eta) = \log \sum_{y} b(y) \exp\left\{\eta^T T(y)\right\}$$

Log Partition Function

Yes! Recall that

$T(y) = y$

$$a(\eta) = \log \sum_y b(y) \exp\left\{\eta^T T(y)\right\}$$

Then, taking derivatives

$E[T(y);\eta]$

$$\nabla_\eta a(\eta) = \frac{\sum_y \boxed{T(y)}\, b(y) \exp\{\eta^T T(y)\}}{\sum_y b(y) \exp\{\eta^T T(y)\}} = \mathbb{E}[T(y); \eta]$$

$P(y;\eta) = b(y) \exp[\eta^T T(y) - a(\eta)]$

Many Other Exponential Models

- There are many canonical exponential family models:
  - Binary $\mapsto$ Bernoulli
  - Multiple Classses $\mapsto$ Multinomial
  - Real $\mapsto$ Gaussian
  - Counts $\mapsto$ Poisson
  - $\mathbb{R}_+ \mapsto$ Gamma, Exponential
  - Distributions $\mapsto$ Dirichlet

Recap

linear

- Linear Regression $h_\theta(x) = \theta^T x$

$\theta^T x$ $\quad$ $1\{\theta^T x\}$ $\quad$ $1(x) = x$

- Logistic Regression $h_\theta(x) = g(\theta^T x)$

$g = \frac{1}{1+e^{-z}}$ $\quad$ $p(x)$

- Multi-class Classification Regression $h_\theta(x) = softmax(\theta_1^T x, \cdots, \theta_k^T x)$

$p(x)$

Recap

- Linear Regression $h_\theta(x) = \theta^T x$ $\quad \underline{\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}}$
- Logistic Regression $h_\theta(x) = g(\theta^T x)$
- Multi-class Classification Regression $h_\theta(x) = softmax(\theta_1^T x, \cdots, \theta_k^T x)$

Recap

- Linear Regression $h_\theta(x) = \theta^T x$ $\quad \theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$
- Logistic Regression $h_\theta(x) = g(\theta^T x)$ $\quad \theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$
- Multi-class Classification Regression $h_\theta(x) = softmax(\theta_1^T x, \cdots, \theta_k^T x)$

Recap

- Linear Regression $h_\theta(x) = \theta^T x$

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

- Logistic Regression $h_\theta(x) = g(\theta^T x)$

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

- Multi-class Classification Regression $h_\theta(x) = softmax(\theta_1^T x, \cdots, \theta_k^T x)$

$$\theta_k := \theta_k + \alpha \sum_{i=1}^{n} (1\{y^{(i)} = k\} - h_\theta(x)_k) x^{(i)}$$

$\vec{\theta}_1 \quad \vec{\theta}_2 \quad \cdots \quad \vec{\theta}_k$

Recap

- Linear Regression $h_\theta(x) = \theta^T x$ $$\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

- Logistic Regression $h_\theta(x) = g(\theta^T x)$ $$\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

- Multi-class Classification Regression $h_\theta(x) = softmax(\theta_1^T x, \cdots, \theta_k^T x)$

$$\theta_k := \theta_k + \alpha \sum_{i=1}^{n} (1\{y^{(i)} = k\} - h_\theta(x)_k)x^{(i)}$$

Is this coincidence?

Generalized Linear Models

We're given features $x \in \mathbb{R}^{d+1}$ and a target y. We want a model.
We first we pick a distribution based on y's type.

Generalized Linear Models

We're given features $x \in \mathbb{R}^{d+1}$ and a target y. We want a model. We first we pick a distribution based on y's type.

- We assume $y \mid x; \theta$ distributed as an exponential family.
  - Binary $\mapsto$ Bernoulli
  - Multiple Classses $\mapsto$ Multinomial
  - Real $\mapsto$ Gaussian
  - Counts $\mapsto$ Poisson
  - $\mathbb{R}_+ \mapsto$ Gamma, Exponential
  - Distributions $\mapsto$ Dirichlet

distribution for prob vectors

K class

[0.1 0.2 0.3 ... 0.4]

Generalized Linear Models

We're given features $x \in \mathbb{R}^{d+1}$ and a target y. We want a model. We first we pick a distribution based on y's type.

- We assume $y \mid x; \theta$ distributed as an exponential family.
  - Binary $\mapsto$ Bernoulli
  - Multiple Classses $\mapsto$ Multinomial
  - Real $\mapsto$ Gaussian
  - Counts $\mapsto$ Poisson
  - $\mathbb{R}_+ \mapsto$ Gamma, Exponential
  - Distributions $\mapsto$ Dirichlet

Neural Network

$\eta = NN(x)$

- Our model is *linear* beacuse we make the natural parameter $\eta = \theta^T x$ in which $\theta, x \in \mathbb{R}^{d+1}$.

Generalized Linear Models

$\int_y y\,p(y)\,dy$

x input y output

inference $h_\theta(x) = \boxed{\mathbb{E}[y \mid x; \theta]}$ is the **output**.

learn $\max_\theta \log p(y \mid x; \theta)$ by maximum likelihood.

$\nabla_\eta a(\eta)$

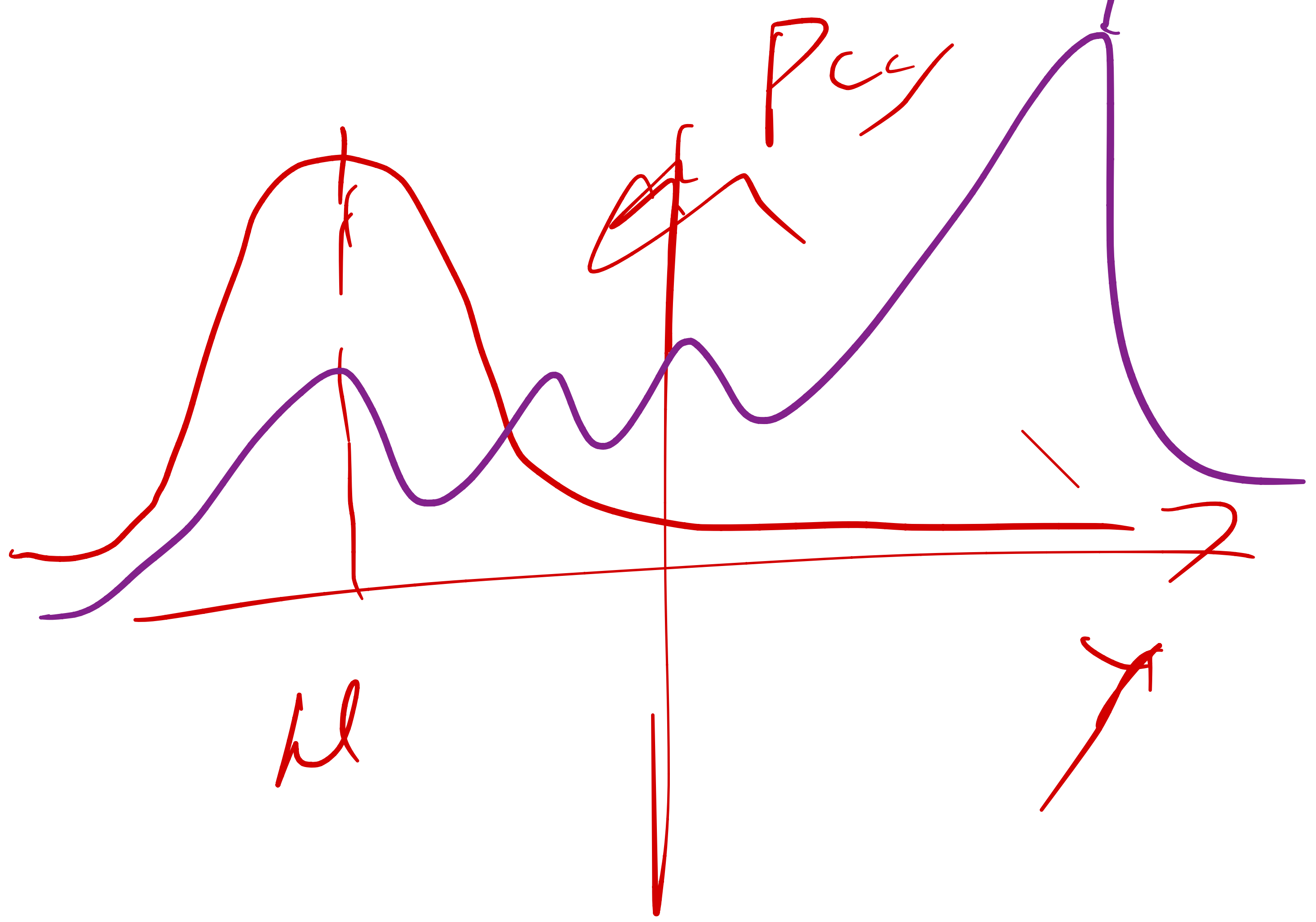
closed-form output

$E(y)$

$f(y) \propto$

$p(y) \propto f(y)$

$p(y) = \frac{f(y)}{Z}$



Generalized Linear Models

inference $h_\theta(x) = \mathbb{E}[y \mid x; \theta]$ is the **output**.

learn $\max_\theta \log p(y \mid x; \theta)$ by maximum likelihood.

$$P(y;\eta) = b(y) \exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

$$a(\eta) = \log \sum_y b(y) \exp\left\{\eta^T T(y)\right\}$$

Then, taking derivatives

$$\nabla_\eta a(\eta) = \frac{\sum_y T(y) b(y) \exp\{\eta^T T(y)\}}{\sum_y b(y) \exp\{\eta^T T(y)\}} = \mathbb{E}[T(y);\eta]$$

Generalized Linear Models

inference $h_\theta(x) = \mathbb{E}[y \mid x; \theta]$ is the **output**.

learn $\max_\theta \log p(y \mid x; \theta)$ by maximum likelihood.

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}.$$

$$a(\eta) = \log \sum_y b(y)\exp\left\{\eta^T T(y)\right\}$$

Then, taking derivatives

$$\nabla_\eta a(\eta) = \frac{\sum_y T(y) b(y) \exp\{\eta^T T(y)\}}{\sum_y b(y) \exp\{\eta^T T(y)\}} = \mathbb{E}[T(y);\eta]$$

$T(y) = y$ for most of the examples you will see in this course

Generalized Linear Models

inference $h_\theta(x) = \mathbb{E}[y \mid x; \theta]$ is the **output**.

learn $\max_\theta \log p(y \mid x; \theta)$ by maximum likelihood.

Generalized Linear Models

inference $h_\theta(x) = \mathbb{E}[y \mid x; \theta]$ is the **output**.

learn $\max_\theta \log p(y \mid x; \theta)$ by maximum likelihood.

algorithm: SGD $\theta^{(t+1)} = \theta^{(t)} + \alpha \left(y^{(i)} - h_{\theta^{(t)}}(x^{(i)}) \right) x^{(i)}$

gradient descent

Constructing GLMs

Constructing GLMs

- Pick an exponential family distribution given the type of y (Possion, Multinomial, Gaussian…)

Constructing GLMs

- Pick an exponential family distribution given the type of y (Possion, Multinomial, Gaussian…)
- $\eta = \theta^T x$, or $\eta_i = \theta_i^T x$ softmax

Constructing GLMs

- Pick an exponential family distribution given the type of y (Possion, Multinomial, Gaussian…)
- $\eta = \theta^T x$, or $\eta_i = \theta_i^T x$
- Training with maximum likelihood estimation

Constructing GLMs

- Pick an exponential family distribution given the type of y (Possion, Multinomial, Gaussian…)
- $\eta = \theta^T x$, or $\eta_i = \theta_i^T x$
- Training with maximum likelihood estimation
- Inference: $h(x) = E[y \mid x]$

Constructing GLMs

- Pick an exponential family distribution given the type of y (Possion, Multinomial, Gaussian…)
- $\eta = \theta^T x$, or $\eta_i = \theta_i^T x$
- Training with maximum likelihood estimation
- Inference: $h(x) = E[y\,|\,x]$

Enjoy closed-form solution for various statistics

easy to sample from

$p(x)$

random. sample.

$x \sim p(x)$

sample

$E(x)$

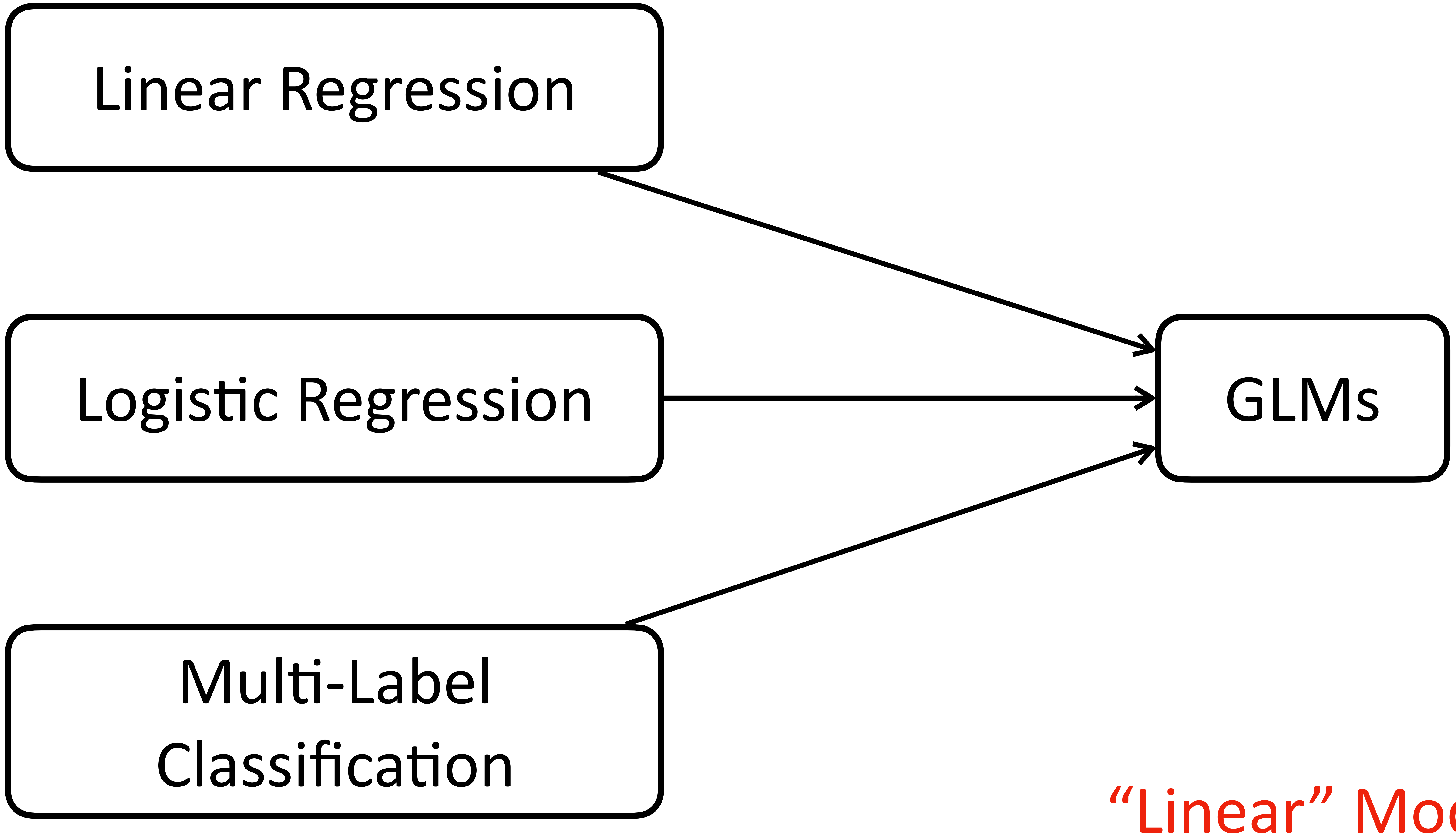
Gaussian

special

$x \sim p(x)$

draw x from $p(x)$

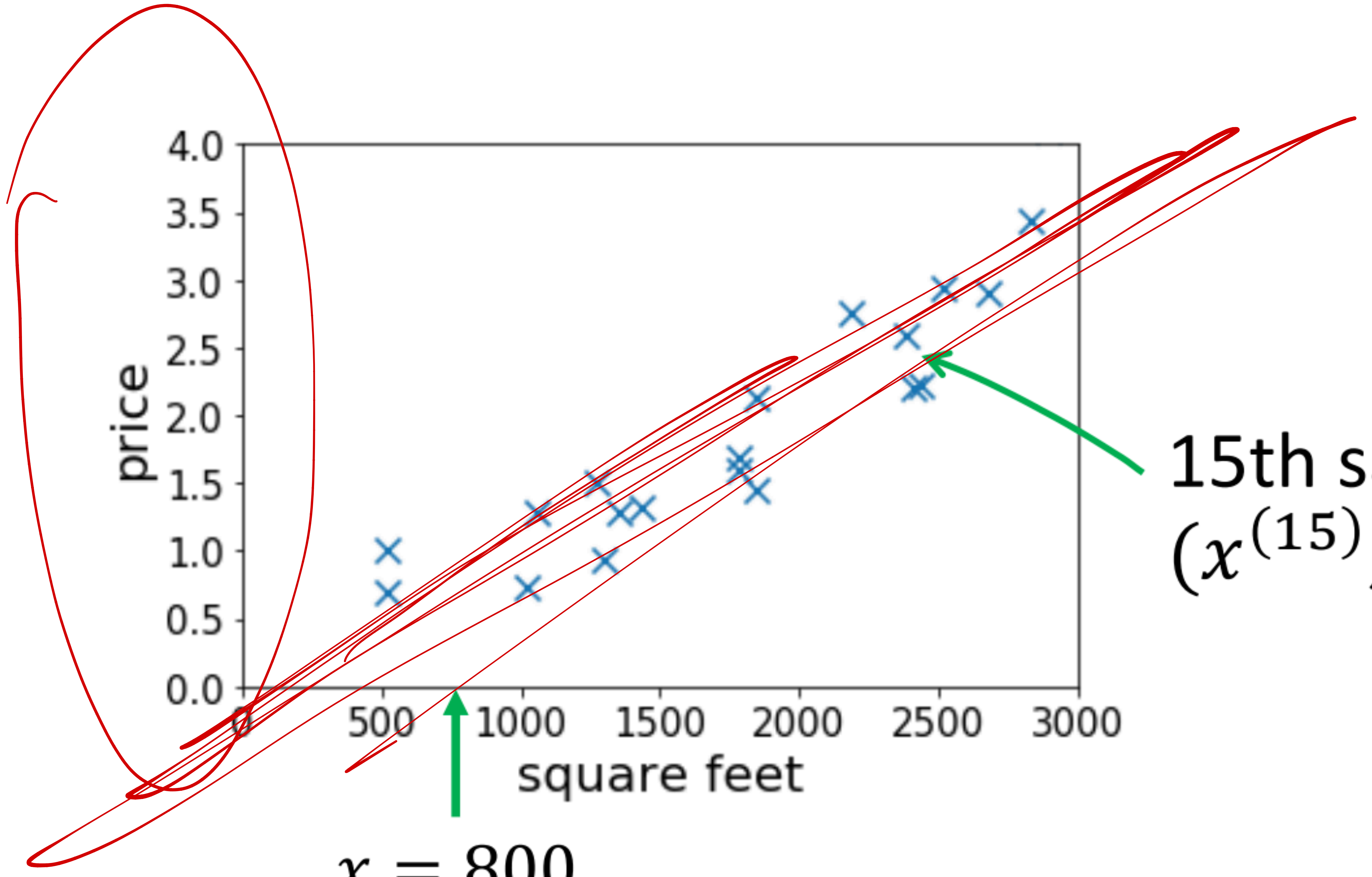
Generalized Linear Models



"Linear" Models

Kernel Methods

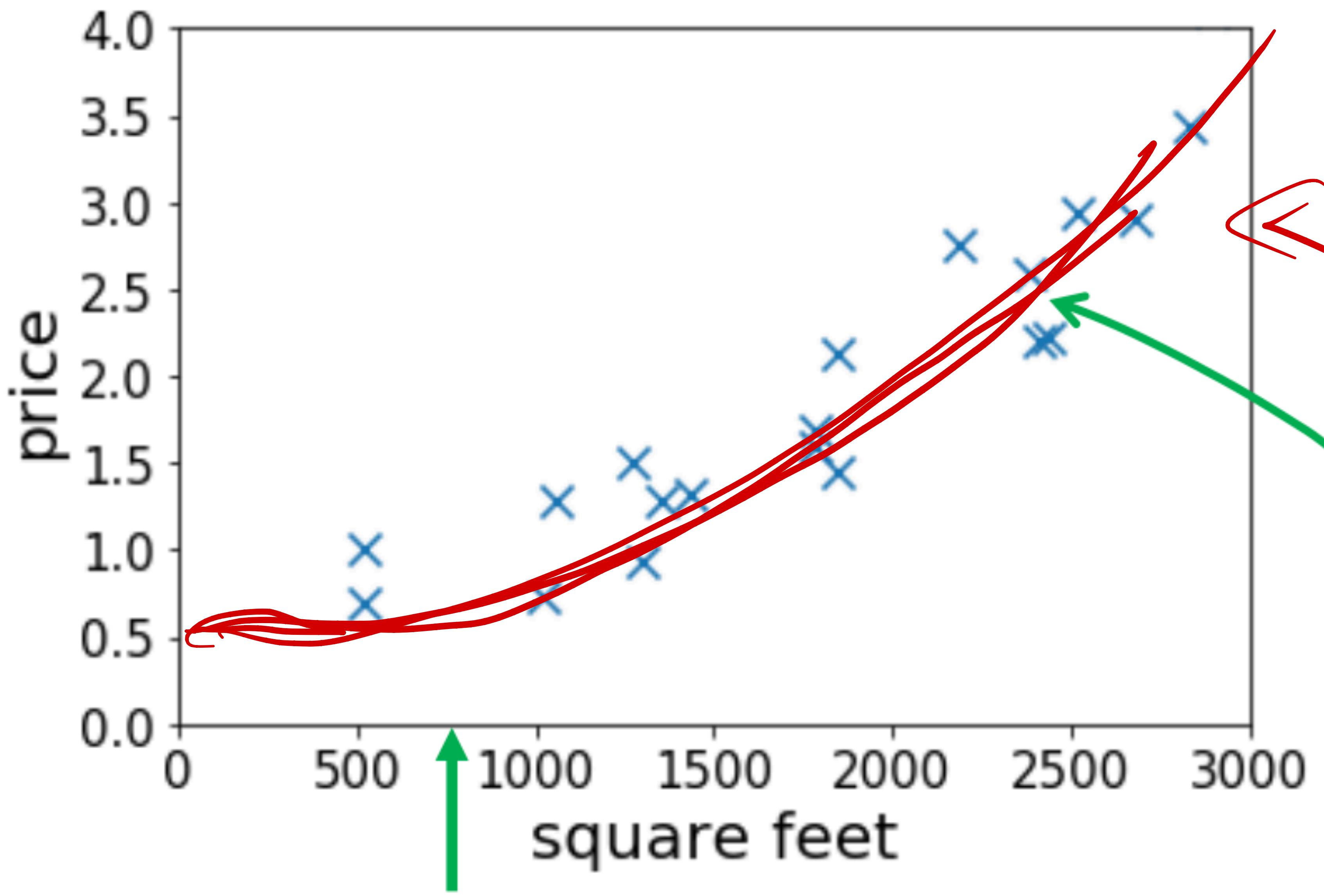
Feature Map



15th sample
$(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

Feature Map

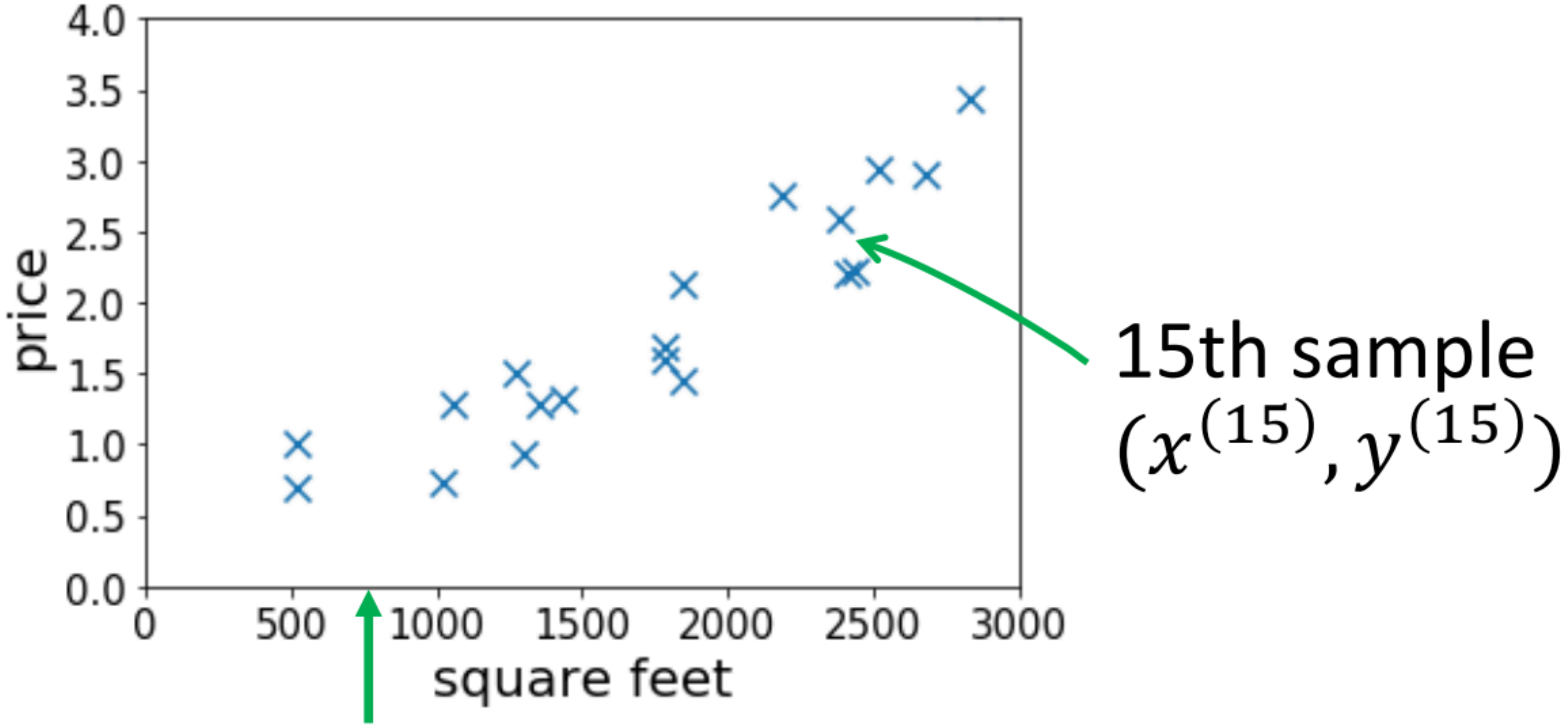


$y = \theta x$

15th sample
$(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

Feature Map

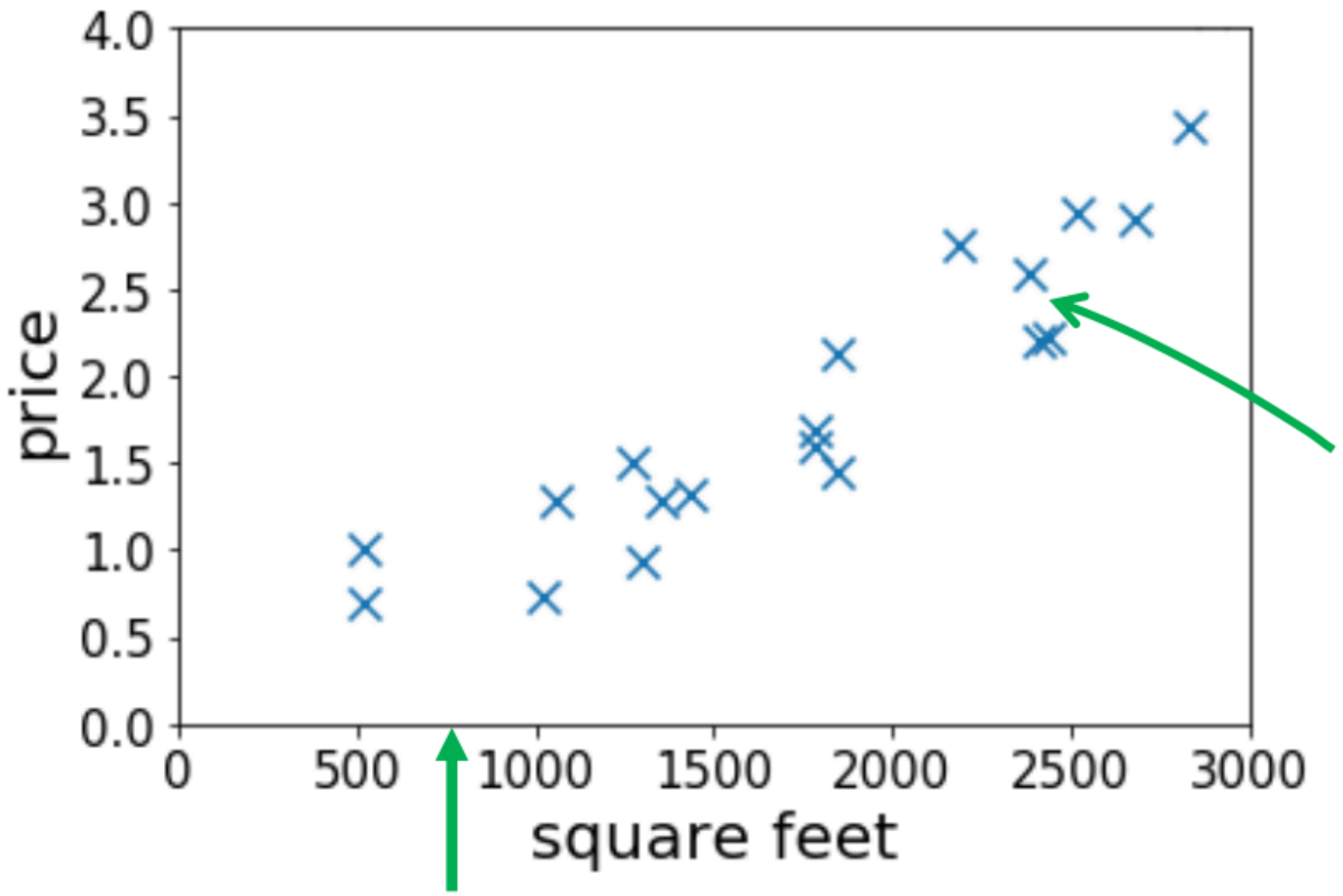


$x = 800$
$y = ?$

$y = \theta x$

$$y = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$$

Feature Map



15th sample
$(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

$$y = \theta x$$

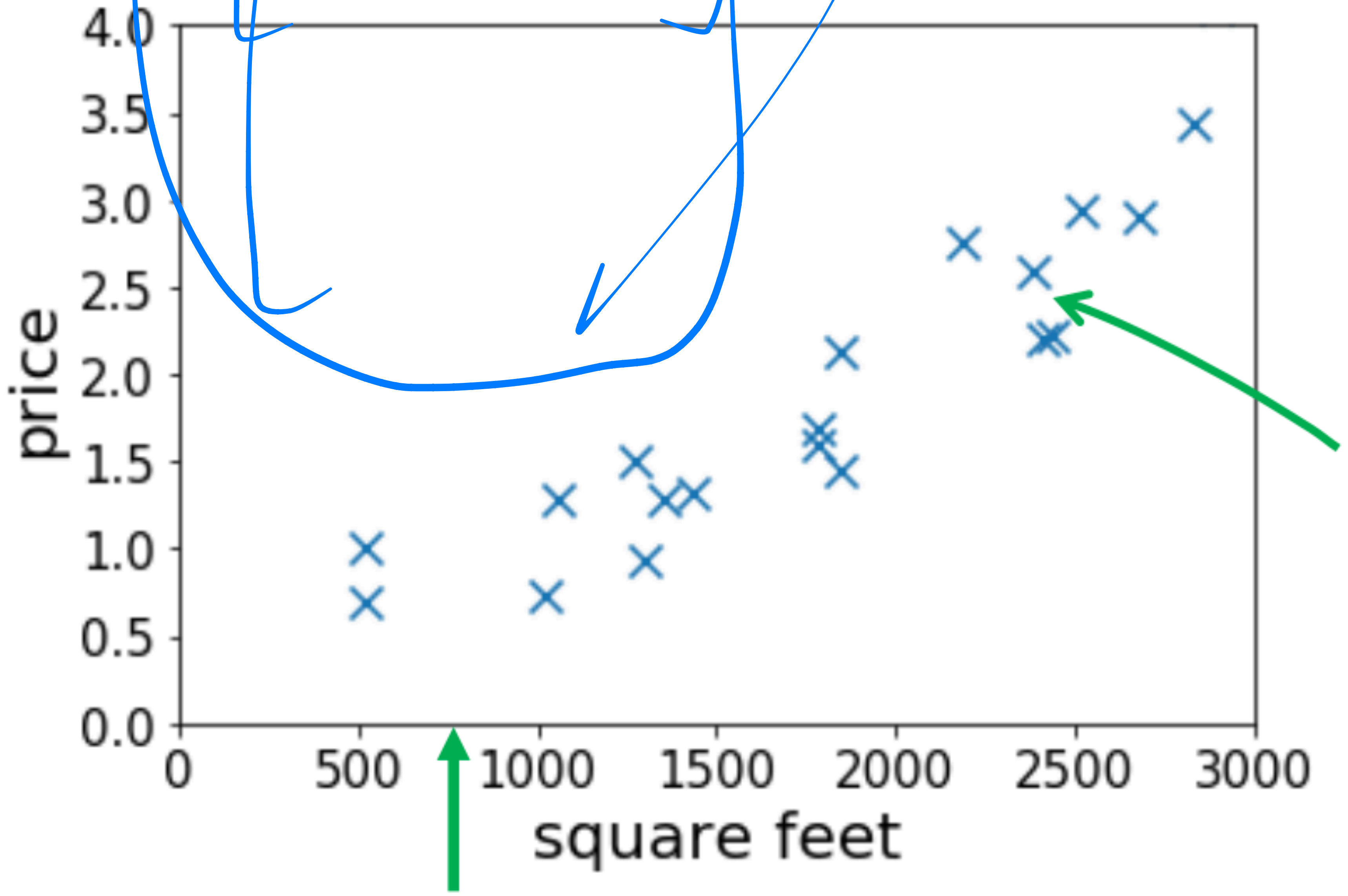
$$y = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \in \mathbb{R}^4.$$

Feature Map

$x \to \begin{bmatrix} e^x \\ \log x \\ x^2+x^3 \end{bmatrix} = \phi(x)$

$\theta^T \phi(x) = y$



15th sample $(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

$$y = \theta x$$

$$y = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \in \mathbb{R}^4.$$

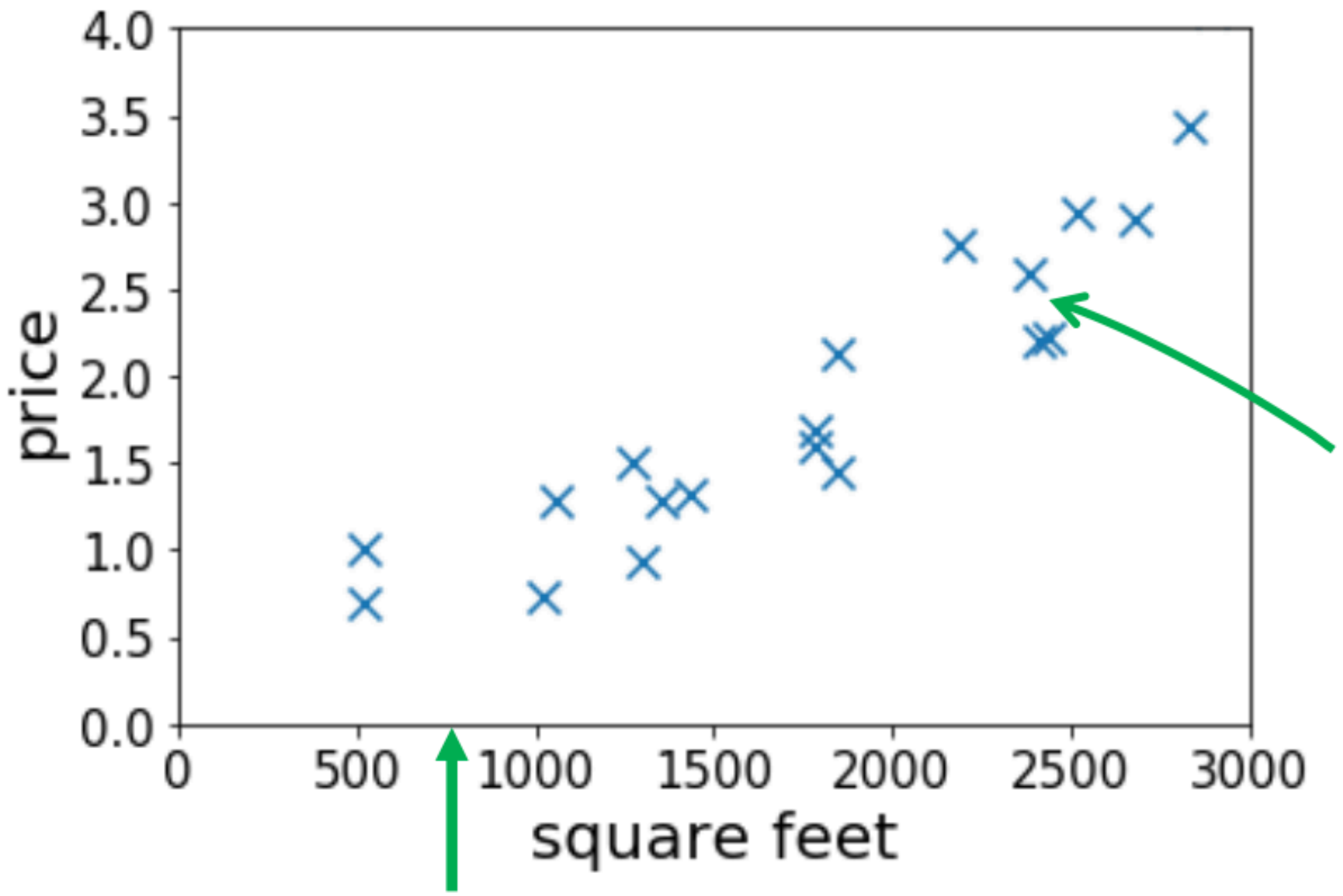
$\theta(\phi)$

Lms

$x \to \phi(x)$

$$y = \theta^T \phi(x)$$

Feature Map



$x = 800$
$y = ?$

15th sample
$(x^{(15)}, y^{(15)})$

$$y = \theta x$$

$$y = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \in \mathbb{R}^4.$$

Feature map
$\phi : R^d \rightarrow R^p$

$\theta \in R^p$

$p >> d$

$$y = \theta^T \phi(x)$$

LMS Update Rule with Features

Linear Regression:

$$\begin{aligned}\theta &:= \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x^{(i)} \\ &:= \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right) x^{(i)}.\end{aligned}$$

With Features:

LMS Update Rule with Features

Linear Regression:

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x^{(i)}$$

$$:= \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)} \right) x^{(i)}.$$

With Features:

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)}) \right) \phi(x^{(i)})$$

LMS Update Rule with Features

Linear Regression:

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x^{(i)}$$

$$:= \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right) x^{(i)}.$$

$\vec{\theta}$

$\theta^T \phi(x^{(i)})$

$O(p))$

With Features:

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

How about Generalized Linear Models with Features?

New Feature Vector Can Be Very High-Dimensional

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_1^2 \\ x_1 x_2 \\ x_1 x_3 \\ \vdots \\ x_2 x_1 \\ \vdots \\ x_1^3 \\ x_1^2 x_2 \\ \vdots \end{bmatrix}$$

Computationally expensive

$x = (x_1, x_2, x_3)$

$x \in R^3$

$\theta^T \phi(x) \quad O(P)$

$\phi(x) \in R^P$

New Feature Vector Can Be Very High-Dimensional

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_1^2 \\ x_1 x_2 \\ x_1 x_3 \\ \vdots \\ x_2 x_1 \\ \vdots \\ x_1^3 \\ x_1^2 x_2 \\ \vdots \end{bmatrix}$$

Computationally expensive

p is large

Is the computation evitable given $\theta \in R^p$?

Kernel Trick

θ is init as 0

θ is θ_0

- If θ is initialized as 0, then at any step of the gradient descent:

anytime

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$$

$\beta_i \in R$

n: # data samples

Kernel Trick

$\theta_0 = 0$ $\qquad \theta_1 = \alpha \sum_{i=1}^{n} \overbrace{(y^{(i)} - \theta^T \phi(x^{(i)}))}^{\beta_i} \phi(x^{(i)})$

- If θ is initialized as 0, then at any step of the gradient descent:

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \underbrace{\left(\beta_i + \alpha \left(y^{(i)} - \theta^T \phi(x^{(i)})\right)\right)}_{\text{new } \beta_i} \phi(x^{(i)})$$

Kernel Trick

- If θ is initialized as 0, then at any step of the gradient descent:

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \underbrace{\left(\beta_i + \alpha \left(y^{(i)} - \theta^T \phi(x^{(i)})\right)\right)}_{\text{new } \beta_i} \phi(x^{(i)})$$

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \theta^T \phi(x^{(i)})\right)$$

Kernel Trick

$\theta_0 = 0$, $\theta_1 = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$, if we prove

we prove
if $\theta_{t-1} = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$
then $\theta_t = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$

- If θ is initialized as 0, then at any step of the gradient descent:

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$\beta_1 \quad \beta_2 \quad \cdots \quad \beta_n$

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \underbrace{\left(\beta_i + \alpha \left(y^{(i)} - \theta^T \phi(x^{(i)})\right)\right)}_{\text{new } \beta_i} \phi(x^{(i)})$$

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \qquad \boxed{\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)})\right)}$$

Kernel Trick

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \boxed{\phi(x^{(j)})^T \phi(x^{(i)})} \right)$$

Kernel Trick

$$\boxed{\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)}$$

Rewrite $\phi(x^{(j)})^T\phi(x^{(i)}) = <\phi(x^{(j)}), \phi(x^{(i)})>$

Kernel Trick

$$\boxed{\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)}$$

Rewrite $\phi(x^{(j)})^T \phi(x^{(i)}) = < \phi(x^{(j)}), \phi(x^{(i)}) >$

$O(p)$

We can precompute all pairwise $< \phi(x^{(j)}), \phi(x^{(i)}) >$ $\in \mathbb{R}$
beforehand, and reuse it for every gradient descent update

Kernel Trick

$$\boxed{\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)}$$

Kernel $K(x, z)$ $\quad \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \mathcal{X}$ is the space of the input

$$K(\vec{x}, \vec{z}) \triangleq \langle \phi(x), \phi(z) \rangle$$

R

The Algorithm

The Algorithm

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j

The Algorithm

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = <\phi(x^{(i)}), \phi(x^{(j)})>$ for all i, j
- Loop $\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$

The Algorithm

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$

Recall that n is the number of data samples

The Algorithm

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = <\phi(x^{(i)}), \phi(x^{(j)})>$ for all i, j
- Loop $$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$$

Recall that n is the number of data samples

Or in vector notation, letting K be the $n \times n$ matrix with $K_{ij} = K(x^{(i)}, x^{(j)})$, we have

$$\beta := \beta + \alpha(\vec{y} - K\beta)$$

Inference

Inference

We do not need to explicitly compute θ !

Inference

We do not need to explicitly compute θ !

$$\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

$$\theta = \sum_{i}^{n} \beta_i \phi(x^{(i)})$$

Inference

We do not need to explicitly compute θ !

$$\theta^T\phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

The Kernel function is all we need for training and inference!

Implicit Feature Map

Do we still need to define feature maps?

$$K(x, z) \triangleq \langle \phi(x), \phi(z) \rangle$$

Implicit Feature Map

Do we still need to define feature maps?

$$K(x, z) \triangleq \langle \phi(x), \phi(z) \rangle$$

What kinds of kernel functions K() can correspond to some feature map ϕ

Example

$\boxed{O(d)}$

d

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

Example

$= \phi(x)^T \phi(z)$

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

Example

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned}
K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\
&= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j)
\end{aligned}$$

Example

$\theta^T x$

$O(d)$

$O(d)$

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

$\phi(z) = \begin{bmatrix} z_1 z_1 \\ z_1 z_2 \\ z_1 z_3 \\ \vdots \\ z_3 z_3 \end{bmatrix}$

Example

$O(d)$ $\phi(x)^T \phi(z)$

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x, z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix} \quad 9$$

Requires O(d^2) compute for feature mapping

Example

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

Requires O(d^2) compute for feature mapping

Requires O(d) compute for Kernel function

Next Lecture

- What kinds of functions would make a kernel function?
- Infinite dimensions of feature mapping?
- Support Vector Machines

Thank You!

Q & A