Minimizing Squared Error (probabilistic interpretation)

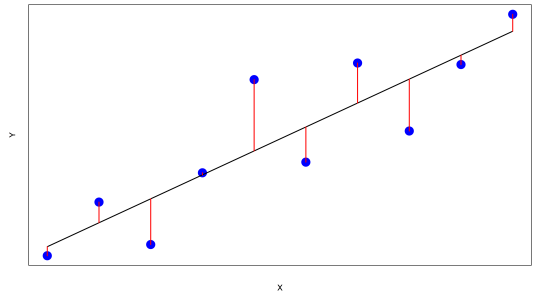
We can write the relationship between input variable x and output variable y as:

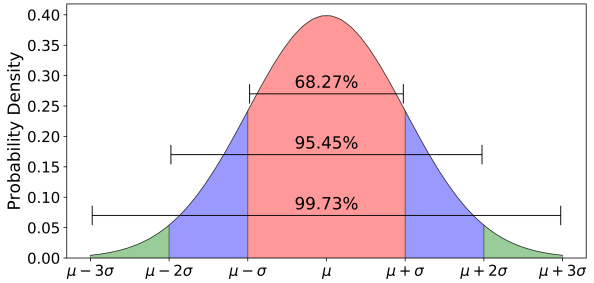
$$y_j = x_j^T \theta + \varepsilon_j$$

And ε_j is an error term which might be unmodeled effect or random noise. Let's assume ε_js are independent and identically distributed ($i.i.d.$) according to a Gaussian distribution:

$$\varepsilon_j \sim N(0, \sigma^2)$$

$$p(\varepsilon_j) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\varepsilon_j^2}{2\sigma^2})$$





Minimizing Squared Error (probabilistic interpretation)

This implies that:

$$p(\varepsilon_j) = p(y_j|x_j;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y_j - x_j^T\theta)^2}{2\sigma^2})$$

So we want to estimate θ such that we maximize the probability of output y given input x over all m training samples:

$\mathcal{L}(\theta) = p(\mathbb{y}|X;\theta)$ (this is called **Likelihood** function)

Minimizing Squared Error (probabilistic interpretation)

Since we assumed independence over ε_j, that implicitly means that our training samples are also independent from each other. Based on this assumption, we can write $\mathcal{L}(\theta)$ as follows:

$$\mathcal{L}(\theta) = \prod_{j=1}^{m} p(y_j | x_j; \theta)$$

$$= \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - x_j^T\theta)^2}{2\sigma^2}\right)$$

Now, we have to estimate θ such that it maximized $\mathcal{L}(\theta)$ to have as high probability as possible. This is called **maximum likelihood**.

Minimizing Squared Error (probabilistic interpretation)

We know (from math) that to find θ that maximized $\mathcal{L}(\theta)$, we can also maximize and strictly increasing function of $\mathcal{L}(\theta)$. In this case it would be easier if we **maximize the log likelihood** $\ell(\theta)$:

$$\ell(\theta) = \log \mathcal{L}(\theta) = \log \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_j - x_j^T\theta)^2}{2\sigma^2}) =$$

$$= m \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{\sigma^2}\frac{1}{2}\sum_{j=1}^{m}(y_j - x_j^T\theta)^2$$

So, maximizing $\ell(\theta)$ is equal to minimizing $\sum_{j=1}^{m}(y_j - x_j^T\theta)^2$

Do you know what is $\sum_{j=1}^{m}(y_j - x_j^T\theta)^2$?

Minimizing Squared Error (probabilistic interpretation)

- This simply shows that under certain assumptions ($\varepsilon_j \sim N(0, \sigma^2)$) and $i.i.d.$) the least-squared regression is equivalent to find maximum likelihood estimate of θ.

- Note that the value of σ^2 does not affect the choice of θ

Linear Regression Assumptions

- **Linearity**: The relationship between x and the mean of y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of x.
- **Independence**: Observations are independent of each other.
- **Normality**: For any fixed value of x, y is normally distributed.

Step back: Statistical Techniques for Data Analysis

Probability vs Statistics: The Difference

Probability versus Statistics

- **Probability**: reasons from populations to samples
  - This is deductive reasoning, and is usually sound (in the logical sense of the word)
- **Statistics**: reasons from samples to populations
  - This is inductive reasoning, and is usually unsound (in the logical sense of the word)

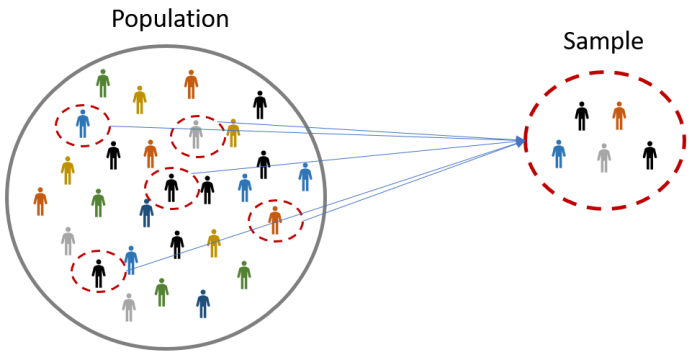
Sampling

Where do the Data come from? (Sampling)

- For groups (populations) that are fairly homogeneous, we do not need to collect a lot of data. (We do not need to sip a cup of tea several times to decide that it is too hot.)
- For populations which have irregularities, we will need to either take measurements of the entire group, or find some way of get a good idea of the population without having to do so
- *Sampling* is a way to draw conclusions about the population without having to measure all of the population. The conclusions need not be completely accurate.
- All this is possible if the sample closely resembles the population about which we are trying to draw some conclusions

What We Want From a Sampling Method

- No systematic bias, or at least no bias that we cannot account for in our calculations
- The chance of obtaining an unrepresentative sample can be calculated. (So, if this chance is high, we can choose not to draw any conclusions.)
- The chance of obtaining an unrepresentative sample decreases with the size of the sample



Estimation

Estimation from a Sample

- In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.
- Estimating some aspect of the population using a sample is a common task. E.g. sample mean are used to estimate population means
- Along with the estimate, we also want to have some idea of the accuracy of the estimate (usually expressed in terms of confidence limits)
- Some measures calculated from the sample are very good estimates of corresponding population values. For example, the sample mean m is a very good estimate of the population mean μ. But this is not always the case. For example, the range of a sample usually underestimates the range of the population

Estimation from a Sample

- We will have to clarify what is meant by a “good estimate”. One meaning is that an estimator is correct on average. For example, on average, the mean of a sample is a good estimator of the mean of the population
- For example, when a number of samples are drawn and the mean of each is found, then average of these means is equal to the population mean
- Such an estimator is said to be *statistically unbiased*

Estimate of the mean and Variance

Mean:

- The arithmetic mean of a sample is $m = \frac{1}{N}\sum_{i=1}^{N} x_i$, where the observations are $x_1, x_2, \ldots, x_N$
- This works very well when the data follow a "normal" distribution
- If we can group the data so that observation x_1 occurs f_1 times, x_2 occurs f_2 times and so on, then the mean is calculated even easier as $m = \frac{1}{N}\sum_i x_i f_i$
- If we define relative frequencies as $p_i = \frac{f_i}{N}$, then the mean is simply the observations weighted by relative frequency. That is $m = \sum_i x_i p_i$
- So, the expected value of a discrete random variable X is:

$$E(X) = \sum_i x_i p(X = x_i)$$

Estimates of the Mean and Variance

Variance: measures how far a set of random numbers are spread out from their average value

- Standard deviation (square root of variance) is estimates as:

$$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2}$$

- Again this is a very good estimate when the data are modeled by Normal distribution.
- For grouped data, this is modified to:

$$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2 f_i}$$

- Again, we can write this in terms of expected values, which estimates the scatter (spread) of values of a random variable X around a mean value $E(X)$:

$$Var(X) = E((X - E(X))^2) = E(X^2) - [E(X)]^2$$

Variance is the "mean of squares minus the squares of means"

Covariance and Correlation

Covariance is a measure of relationship between two random variables:

$$cov(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N-1} = \frac{(\sum_i x_i y_i) - N\bar{x}\bar{y}}{N-1}$$

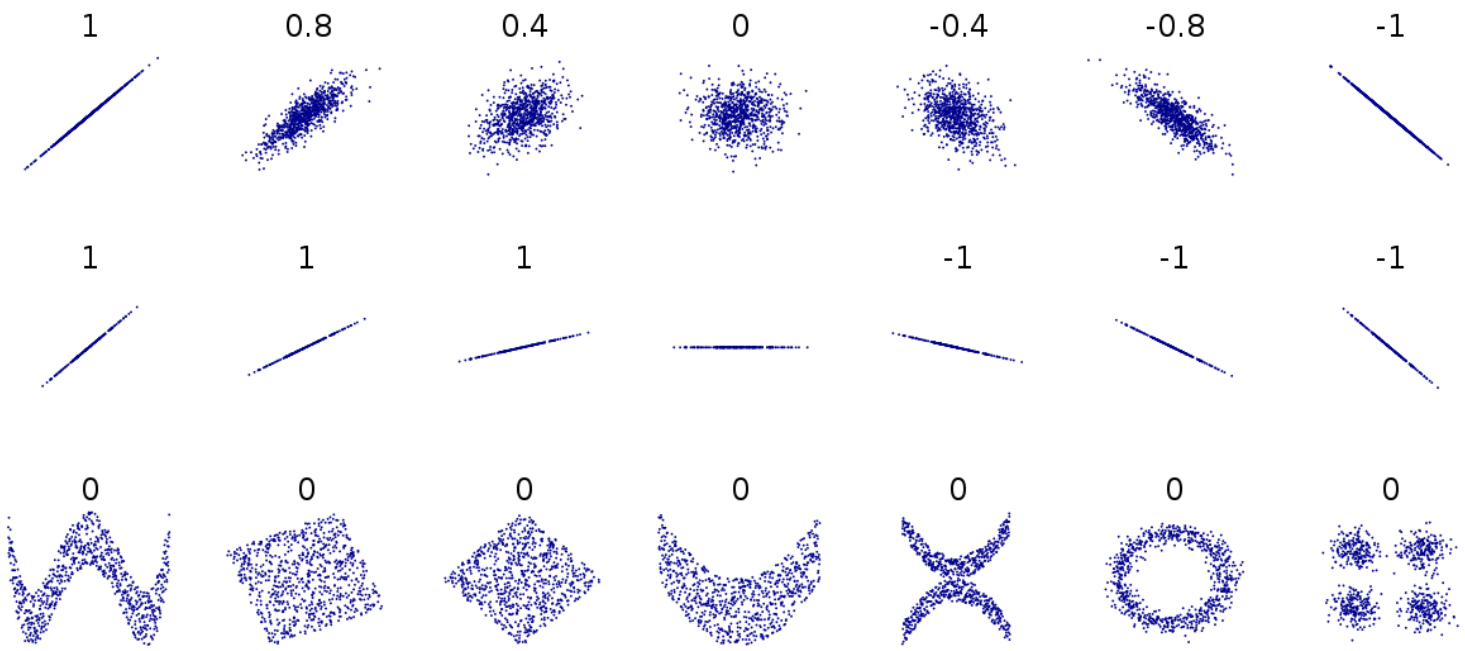
Correlation is a measure to show how strongly a pair of random variables are related

- The formula for computing correlation between x and y is:

$$r = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

This is also called *Pearson's correlation coefficient*

Correlation



Pearson correlation in some sets of (x,y). [Wikipedia]

Correlation

- The correlation coefficient is a number between -1 and $+1$ that shows whether a pair of variable x and y are associated or not and whether their scatter in the association is high or low:
  - A value close to 1 shows that high values of x are associated with high values of y and low values of x are associates with low values of y, and scatter is low
  - A value near 0 indicates that there is no association and there is a large scatter
  - A value close to -1 suggests a strong inverse association between x and y
- Correlation is only appropriate when x and y are roughly linearly associated (does not work well when the association is curved)

What Does Correlation Mean?

- r is a quick way of checking whether there is some linear association between x and y
- The sign of the value tells you the direction of the association
- All that the numerical value tells you is about the scatter in the data
- The correlation coefficient does not model any relationship. That is, given a particular x you cannot use the r value to calculate a y value
  - It is possible for two datasets to have the same correlation, but different relationships
  - It is possible for two datasets to have different correlations but the same relationship
- MORAL: Do not use correlations to compare datasets. All you can derive is whether there is a positive or negative relationship between x and y
- ANOTHER MORAL: Do not use correlation to imply x causes y or the other way around

Regression

Univariate linear regression: (only one independent variable)

Goal: fit a line such that $\hat{y} = \theta_0 + \theta_1 x$

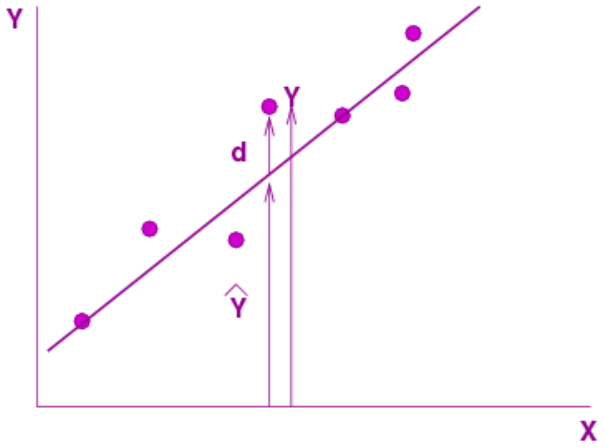
How: minimize $J(\theta) = \sum_j (y_j - \hat{y}_j)^2$ (Least squares estimator)

- If we expand the explicit solution we saw before, we can see:

$$\theta_1 = \frac{cov(x, y)}{var(x)}$$

Where $cov(x, y)$ is the covariance of x and y

- and $\theta_0 = \bar{y} - \theta_1 \bar{x}$



Meaning of the coefficients

- θ_1: change in y that accompanies a unit change in x
- If the values of x were assigned at random, then θ_1 estimates the unit change in y caused by a unit change in x
- If the values of x were not assigned at random (for example, they were data somebody observed), then the change in y will include the change in x and any other confounding variables that may have changed as a result of changing x by 1 unit. So, you cannot say for example, that a change of x by 1 unit causes θ_1 units of change in y
- $\theta_1 = 0$ means there is no linear relationship between x and y , and then best we can do is simply say is $\theta_0 = \bar{y}$. Estimating the sample mean is therefore a special case of the MSE criterion.

Finding the parameters for univariate linear regression

Univariate linear regression

In univariate regression we aim to find the relationship between y and one independent variable x.

Example:

Suppose we want to investigate the relationship between people's height and weight. We collect m height and weight measurements:

$$(h_j, w_j), \qquad j = 1, \ldots, m$$

Univariate linear regression assumes a linear relation $w = \theta_0 + \theta_1 h$, with parameters θ_0, θ_1 chosen such that the sum of squared residuals $\sum_{j=1}^{m} (w_j - \theta_0 + \theta_1 h_j)^2$ is minimized

Univariate linear regression

In order to find the parameters we take partial derivatives, set the partial derivatives to 0 and solve for θ_0. As we saw before, this will lead to:

$$\theta_1 = \frac{cov(h,w)}{var(h)} = \frac{\sum_{j=1}^{m}(h_j - \bar{h})(w_j - \bar{w})}{\sum_{j=1}^{m}(h_j - \bar{h})^2}$$

$$\theta_0 = \bar{w} - \theta_1 \bar{h}$$

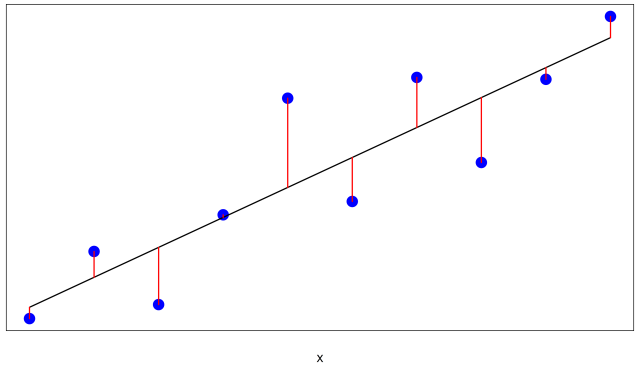
Linear regression: intuitions

- Adding a constant to all x-values (a translation) will affect only the intercept but not the regression coefficient (since it is defined in terms of deviations from the mean, which are unaffected by a translation).
- So we could zero-centre the x -values by subtracting x, in which case the intercept is equal to y .
- We could even subtract y from all y -values to achieve a zero intercept, without changing the problem in an essential way.

Linear regression: intuitions

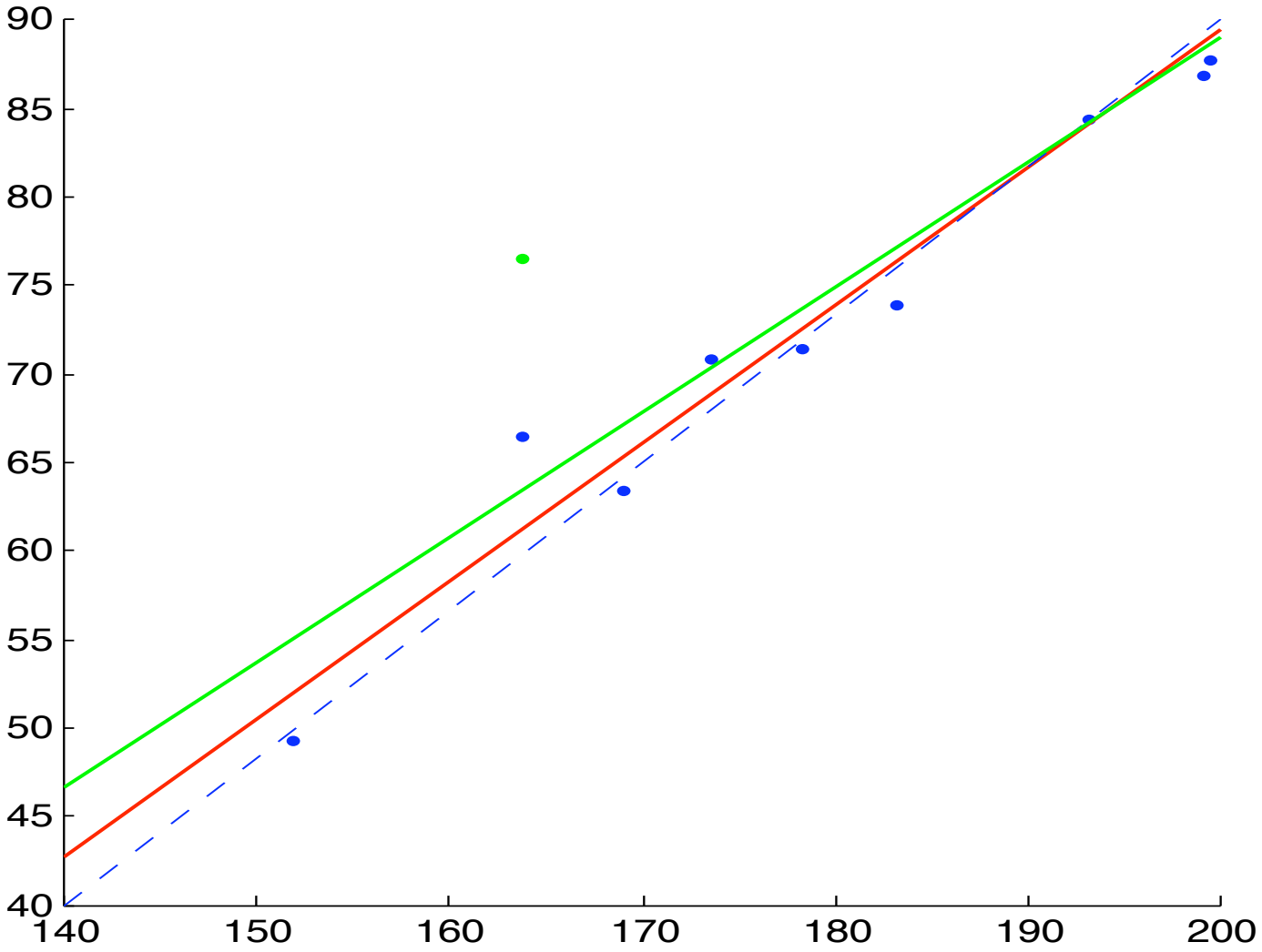
Another important point to note is that the **sum of the residuals** of the least square **is zero**

$$\sum_{j=1}^{m} \left(y_j - x_j^T \theta\right) = 0$$



While this property is intuitively appealing it is worth keeping in mind that it also makes linear regression susceptible to *outliers:* points that are far removed from the regression line, often because of the measurement error.

The effect of outliers



The effect of outliers

Shown on previous slide:

- Suppose that, as the result of a transcription error, one of the weight values from the previous example of univariate regression is increased by 10 kg. The diagram shows that this has a considerable effect on the least-squares regression line.
- Specifically, we see that one of the blue points got moved up 10 units to the green point, changing the red regression line to the green line.

Multiple linear regression

Multiple Regression

Often we are interested in modelling the relationship of y to several other variables. In observational studies the value of y may be affected by the value of several variables.

Example:

Suppose we want to predict people's weight from their weight and body fame size (usually measured by wrist size). We collect m height, weight and body frame measurement:

$$(h_j, w_j, f_j), \qquad j = 1, \ldots, m$$

Multiple Regression

Similar to before the linear regression model is:

$$\hat{w} = \theta_0 + \theta_1 h + \theta_2 f,$$

And similar to univariate linear regression, we have to minimize the sum of squared residuals

$$\sum_{j=1}^{m} \left(w_j - (\theta_0 + \theta_1 h_j + \theta_2 f_j)\right)^2$$

- Including more variables can give a narrower confidence interval on the prediction being made
- With many variables, the regression equations and the expressions for the θ are expressed better using a matrix representation(as we used before) for sets of equations

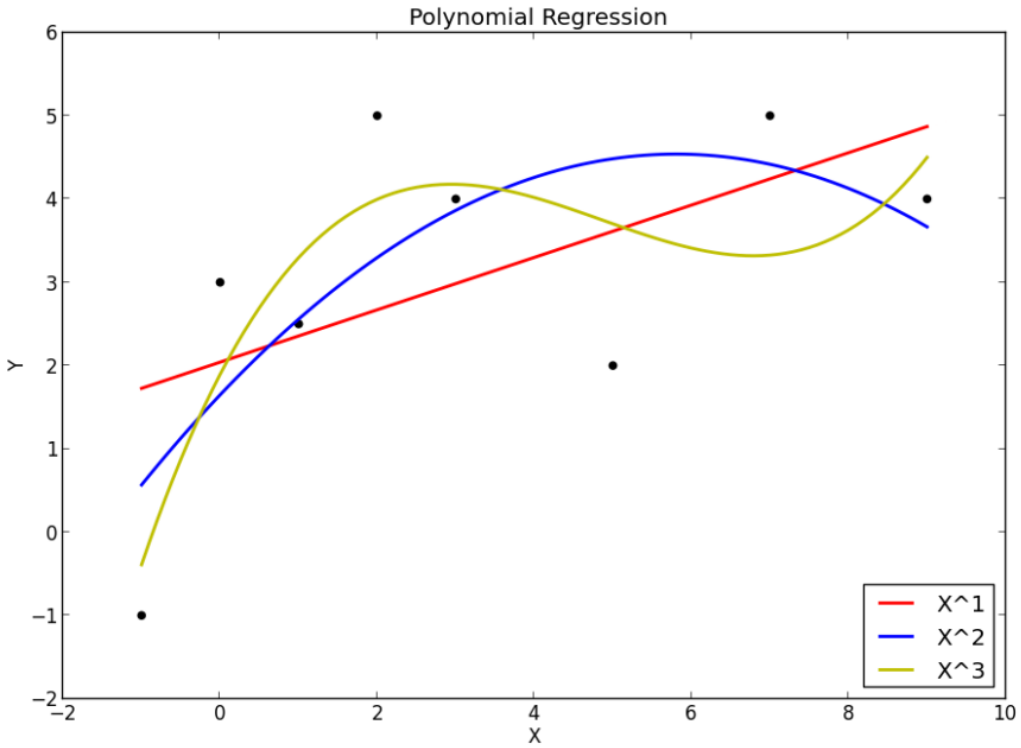
Linear Regression for curve shapes

You may think that linear regression produces straight lines or plains and nonlinear equations models produce curvature!!

Well that's not completely correct. With some tricks we can produce curves with linear regression

Linear Regression for curve shapes

Example



Linear Regression for curve shapes

- In this example we can predict the output with different models:
- $\hat{y} = \theta_0 + \theta_1 x_1$
- $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2,\ x_2 = x_1^2. \rightarrow\ \hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$
- $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^3, x_2 = x_1^2,\ x_3 = x_1^3 \rightarrow\ \hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

As you can see, these nonlinear models can still be treated like linear regression and they can fit curvature. They are still linear in parameters. (Nonlinear regression is not linear in parameters, e.g. $y = \frac{\theta_1 x}{\theta_2 + x}$)

Linear Regression for curve shapes

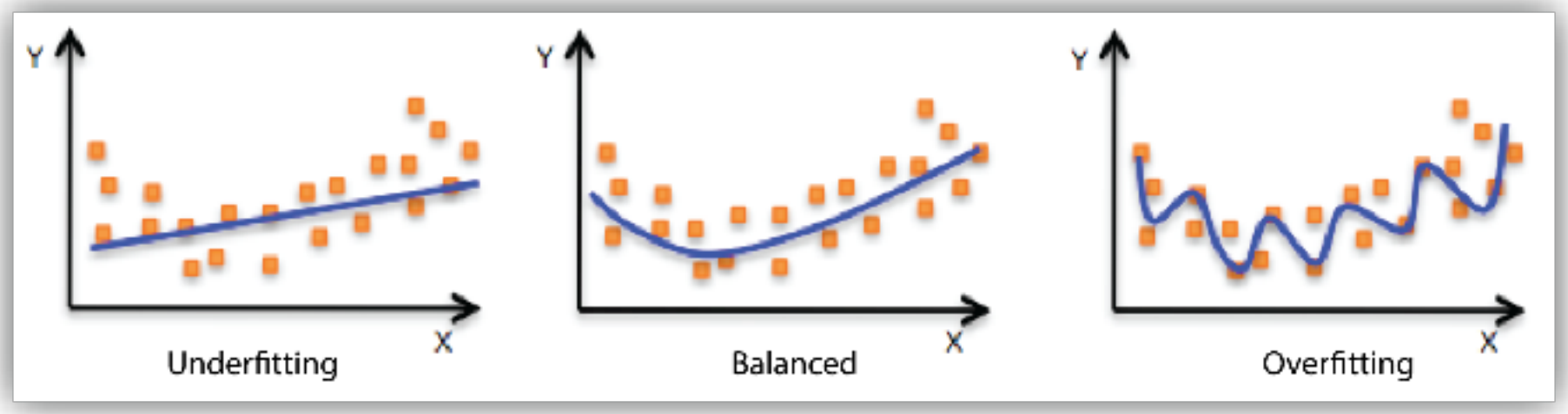


Image from AWS website

Linear Regression for curve shapes

Question:

How to control for degree of complexity of the model to avoid overfitting?

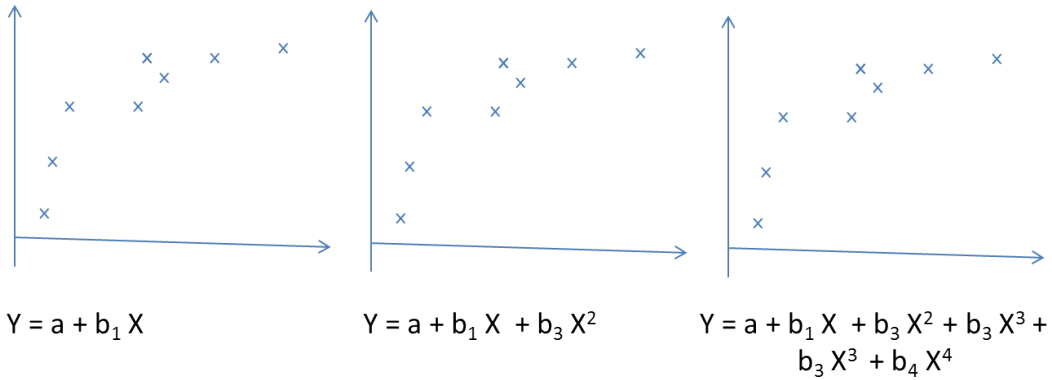
Regularisation

Regularisation

Regularisation is a general method to avoid overfitting by applying additional constraints to the weight vector. A common approach is to make sure the weights are, on average, small in magnitude: this is referred to as *shrinkage*.

Recall the setting for regression in terms of cost minimization.

- Can add penalty terms to a cost function, forcing coefficients to shrink to zero



Regularisation

- MSE as a cost function, given data $(x_1, y_1), \ldots, (x_m, y_m)$

$$J(\theta) = \sum_j (y_j - h_\theta(x_j))^2 + \lambda \sum_i \theta_i^2$$

- Parameter estimation by optimisation will attempt to values for $\theta_0, \ldots, \theta_n$ s.t. $J(\theta)$ is a minimum

- Similar to before, this can be solved by gradient descent or take the derivatives and set them to zero

Regularisation

The multiple least-squares regression problem is an optimisation problem, as we saw, and can be written as:

$$\theta^* = \arg\min_{\theta}(y - X\theta)^T(y - X\theta)$$

The regularised version of this is then as follows:

$$\theta^* = \arg\min_{\theta}((y - X\theta)^T(y - X\theta) + \lambda||\theta||^2$$

Where $||\theta||^2 = \sum_i \theta_i^2$ is the squared norm of the vector θ, or equivalently, the dot product $\theta^T\theta$; λ is a scalar determining the amount of regularisation.

Regularisation

This regularised problem still has a closed-form solution:

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

where I denotes the identity matrix. Regularisation amounts to adding λ to the diagonal of $X^T X$, a well-known trick to improve the numerical stability of matrix inversion. This form of least-squares regression is known as ***ridge regression***.

An interesting alternative form of regularised regression is provided by the ***lasso***, which stands for 'least absolute shrinkage and selection operator'. It replaces the ridge regularisation term $\sum_i \theta_i^2$ with the sum of absolute weights $\sum_i |\theta_i|$. The result is that some weights are shrunk, but others are set to 0, and so the *lasso regression favours sparse solutions*.

Train, Validation & Test Data

- **Train Data:** the data that we use to learn our model and its parameters
- **Validation Data**: The data (unseen by the model) used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
- **Test Data:** unseen data by the model that we use to test the model and shows how well our model generalizes. In practice, we don't know the ground truth (label) for this data

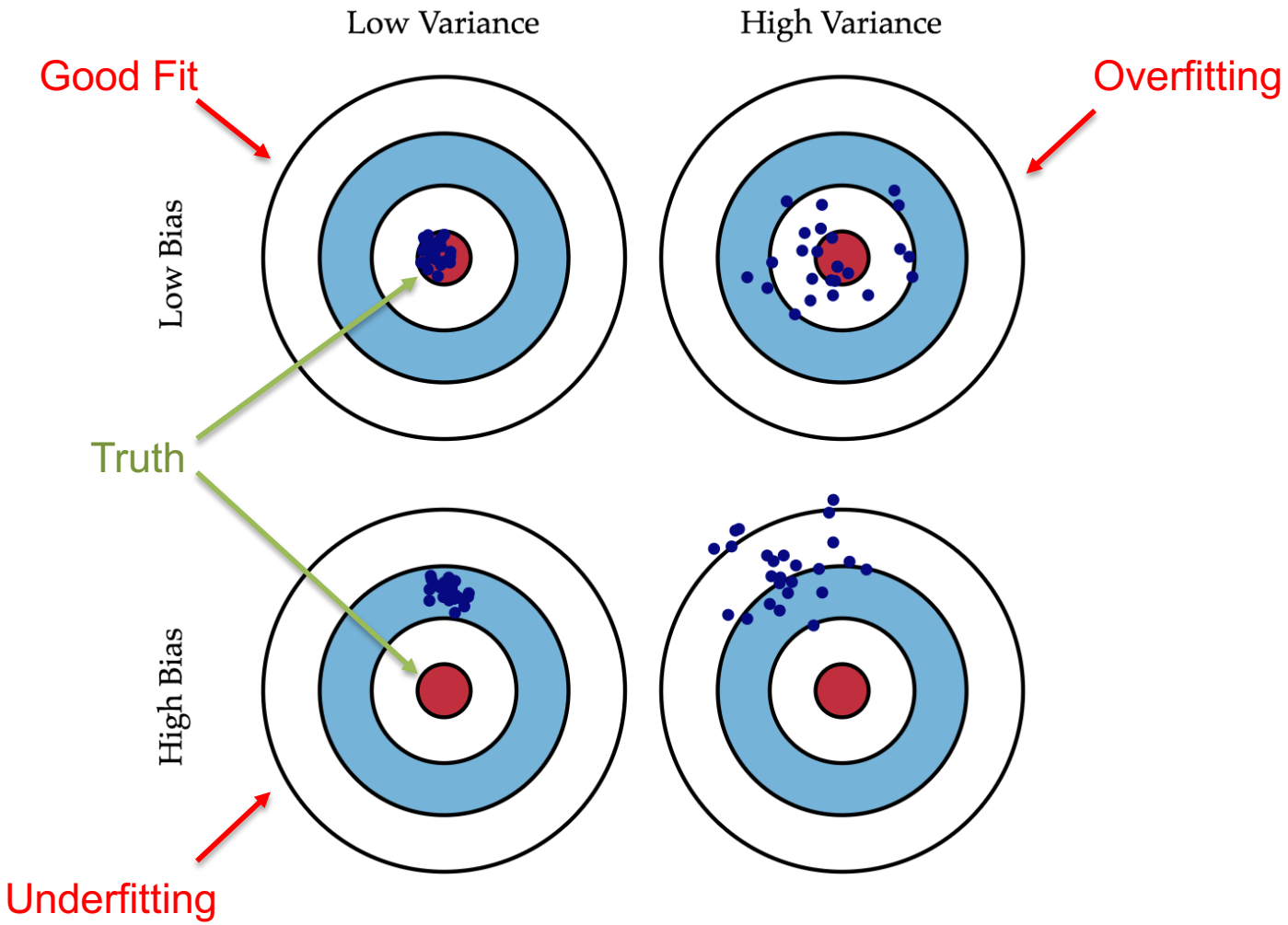
Train, Validation & Test Data

Ideally, we would like, to get the same performance on test set as we get on validation/training set. This is called *Generalization*.

Generalization is the model ability to adapt properly to new, previously unseen data drawn from the same distribution as the one used to create the model.

Bias-Variance Tradeoff

Bias-Variance Tradeoff



Source: Scott-Fortmann,Understanding Bias-variance tradeoff

Bias-Variance Tradeoff

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

Bias-Variance Tradeoff

- In supervised learning, **underfitting** happens when a model is unable to capture the underlying pattern of the data. These models usually have high bias and low variance.
- In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data.

Bias-Variance Decomposition

It can be shown that, when we assume $y = f + \epsilon$ and we estimate f, with $\hat{f}$, then the expectation of error:

$$E[(y - \hat{f})^2] = (f - E[\hat{f}])^2 + Var(\hat{f}) + Var(y)$$

So., the mean of squared error (MSE) can be written as:

$$MSE = Bias^2 + Variance + irreducable\ error$$

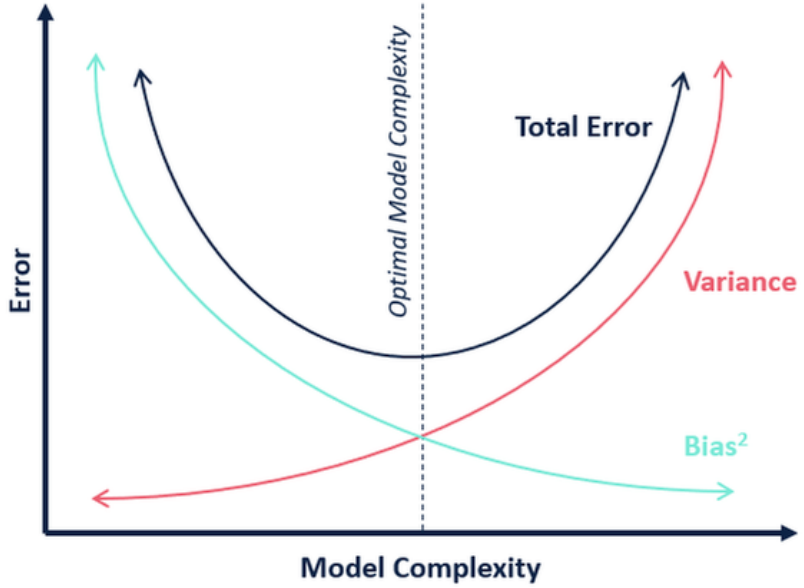
- **Irreducible error** or inherent uncertainty is associated with a natural variability in a system (noise). It can not be not reduced since it is due to unknown/unpredictable factors or simply due to chance.
- **Reducible error**, as the name suggests, can be and should be minimized further by adjustments to the model.

Bias-Variance Tradeoff

- When comparing unbiased estimators, we would like to select the one with minimum variance
- In general, we would be comparing estimators that have some bias and some variance
- If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.
- We have to choose a complexity that makes a good tradeoff between bias and variance.

Bias-Variance Tradeoff

- Often, the is to find the balance between bias and variance such that we minimize the overall (total) error.
- There is not a quantitative way to find this balanced error point. Instead, you will need to leverage (preferably on an unseen validation data set) and adjust your model's complexity until you find the iteration that minimizes overall error.



Model Evaluation

Model Evaluation

The most popular metrics are:

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{y}_j)^2}$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{m}\sum_{j=1}^{m}|(y_j - \hat{y}_j)|$$

- R-squared ($[-\infty, 1]$)

$$R^2 = 1 - \frac{\sum_{j=1}^{m}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{m}(y_j - \bar{y}_j)^2}$$

- Adjuster R-squared

$$R^2_{adjusted} = 1 - [\frac{(1 - R^2)(m-1)}{m-n-1}]$$

Where m is the total number of samples and n is the number of predictors/features.

- R-squared represents the portion of variance in the output that has been explained by the model

Model Evaluation

Example:

Case 1		Case 2			Case 3		
Var1	Y	Var1	Var2	Y	Var1	Var2	Y
x1	y1	x1	2*x1	y1	x1	2*x1+0.1	y1
x2	y2	x2	2*x2	y2	x2	2*x2	y2
x3	y3	x3	2*x3	y3	x3	2*x3 + 0.1	y3
x4	y4	x4	2*x4	y4	x4	2*x4	y4
x5	y5	x5	2*x5	y5	x5	2*x5 + 0.1	y5

	Case 1	Case 2	Case 3
R_squared	0.985	0.985	0.987
Adj_R_squared	0.981	0.971	0.975

- As can be seen adjusted R-squared does a better job at penalizing case 2 & 3 for having more variables without adding any extra information

Model Evaluation

- the absolute value of RMSE does not actually tell how bad a model is. It can only be used to compare across two models
- Adjusted R-squared easily tells about the quality of the model as well. For example, if a model has adjusted R-squared equal to 0.05 then it is definitely poor.
- However, if you care only about prediction accuracy then RMSE is best. It is computationally simple, easily differentiable and present as default metric for most of the models.

Some further issues in learning linear regression models

Categorical Variables

- "Indicator" variables are those that take on the values 0 or 1.
- They are used to include the effect of categorical variables, for example, if d is a variable that takes value 1 if a patient takes a drug and 0 if the patient does not. Suppose we want the effect of drug d on blood pressure y, keeping age x constant

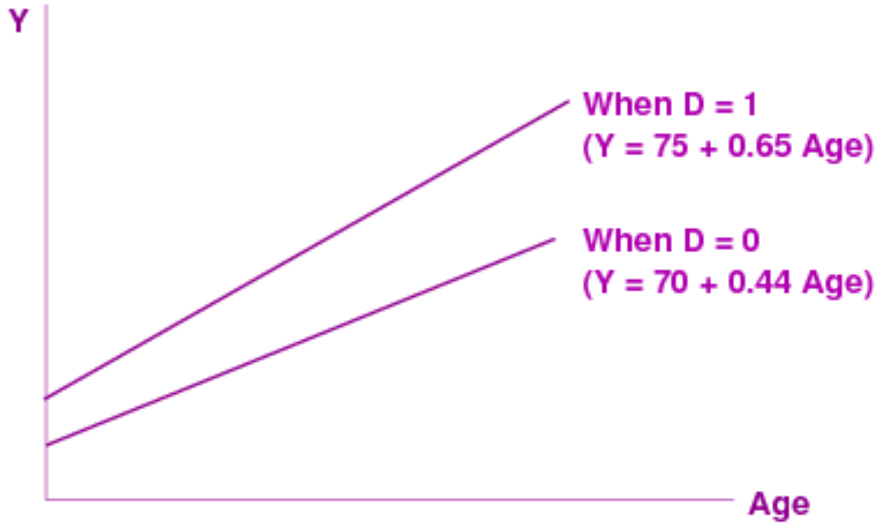
$$y = 70 + 5d + 0.44x$$

- So taking drug d (a unit change in d) makes a difference of 5 units in blood pressure, provided age is held constant.

Categoric Variables

How do we capture any interaction effect between age and drug intake.
Introduce a new indicator variable $z = d \times x$

$$y = 70 + 5d + 0.44x + 0.21z$$



Model Selection

Suppose there are a lot of variables/features (x), some of which may be representing products, powers, etc.

- Taking all the features will lead to an overly complex model. There are 3 ways to reduce complexity:
  1. Subset-selection, by search over subset lattice. Each subset results in a new model, and the problem is one of model-selection
  2. Shrinkage, or regularization of coefficients to zero, by optimization. There is a single model, and unimportant variables have near-zero coefficients.
  3. Dimensionality-reduction, by projecting points into a lower dimensional space (this is different to subset-selection, and we will look at it later)

Model Selection

Subset-selection: we want to find a subset of variables/features which performs well and get rid of redundant features. This is also called stepwise regression.

- Historically, model-selection for regression has been done using "forward-selection", "backward-elimination", or "bidirectional" methods
  - These are greedy search techniques that either:
    - (a) start with no variable and at each step add one variable whose addition gives the most improvement to the fit
    - (b) start with all variables and at each step remove one whose loss gives the most insignificant deterioration of the model fit;
    - (c) a combination of the above, testing at each step for variables to be included or excluded.

Model Selection

- This greedy selection done on the basis of calculating the fit quality (often using R-squared, which denotes the proportion of total variation in the dependent variable Y that is explained by the model)
- Given a model formed with a subset of variables X, it is possible to compute the observed change in R-squared due to the addition or deletion of some variable x
- This is used to select greedily the next best move in the graph-search

To set other hyper-parameters, such as shrinkage parameter in regularisation, we can use grid search

Local (nearest-neighbor) regression

Local Learning

- Related to the simplest form of learning: rote learning, or memorization
- Training instances are searched for instance that most closely resembles query or test instance
- The instances themselves represent the knowledge
- Called: nearest-neighbor, instance-based, memory-based or case-based learning; all forms of local learning
- The similarity or distance function defines “learning”, i.e., how to go beyond simple memorization
- Intuition — classify an instance similarly to examples “close by” — neighbors or exemplars
- A form of lazy learning – don’t need to build a model!

Nearest neighbor for numeric prediction

- Store all training examples $(f(x_i), x_i)$

Nearest neighbour

- Given query instance x_q,
- First locate nearest neighbour x_n,
- Then estimate $\hat{y}_q = f(x_n)$

K-nearest neighbour

- Given x_q, take mean of f values of k nearest neighbour

$$\hat{y}_q = \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

Distance function

The distance function defines what is "learned", i.e., predicted

- Most commonly distance function used is Euclidean distance
- If instant x_i is defined with n feature values

$$\langle x_{i1}, \ldots, x_{in} \rangle$$

- The distance between two instances x_i, x_j is defined as

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

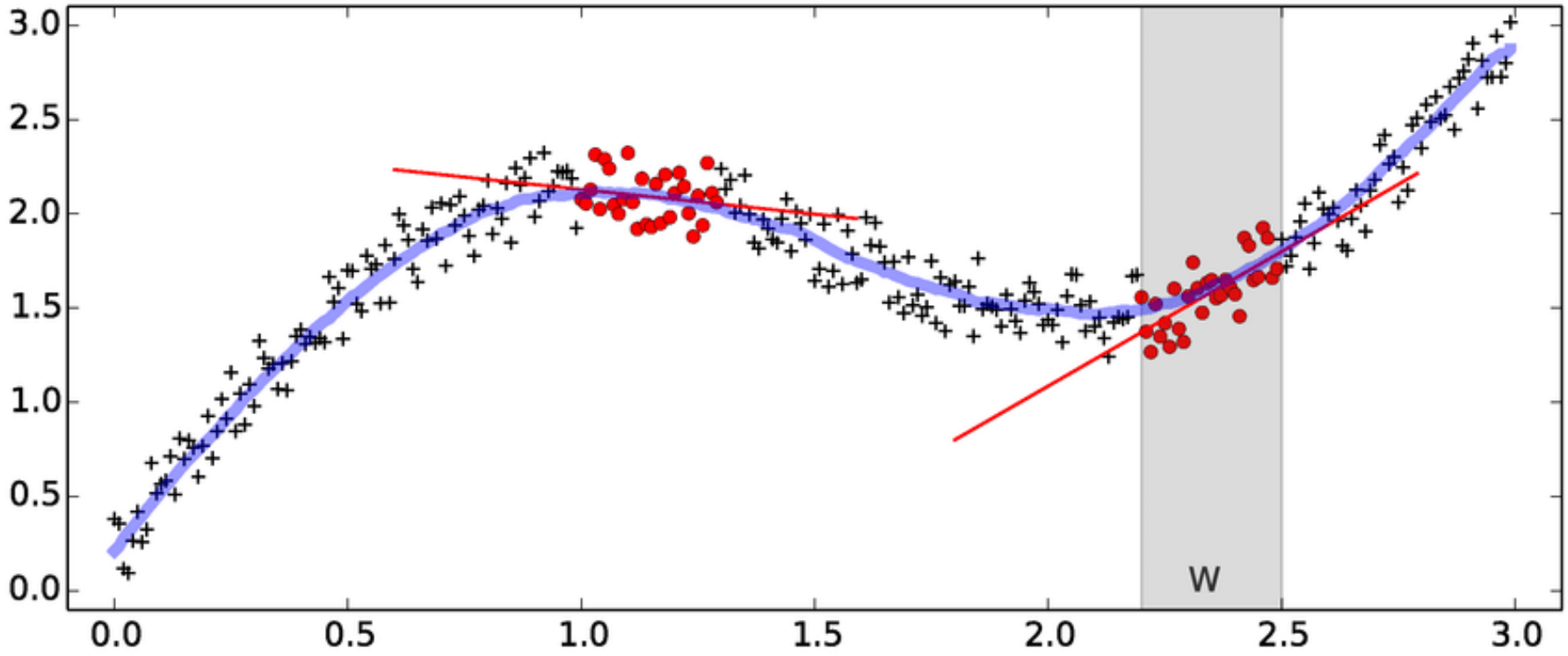
Local regression

Use KNN to form a local approximation to f for each query point x_q using a linear function of the form:

$$\hat{y} = \hat{f}(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

- Fit the liner function to k nearest neighbor of the query point
- Linear, quadratic or higher-order polynomial can be used
- Produces “piecewise approximation” so can be useful for non-linear functions, or data with changing distributions

Local regression



[From, Locally-weighted homographies for calibration of imaging systems, by P. Ranganathan & E. Olson]

Acknowledgements

- Material derived from slides for the book
  "Elements of Statistical Learning (2nd Ed.)" by T. Hastie, R. Tibshirani & J. Friedman. Springer (2009) http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Material derived from slides for the book
  "Machine Learning: A Probabilistic Perspective" by P. Murphy MIT Press (2012)
  http://www.cs.ubc.ca/~murphyk/MLbook
- Material derived from slides for the book "Machine Learning" by P. Flach Cambridge University Press (2012) http://cs.bris.ac.uk/~flach/mlbook
- Material derived from slides for the book
  "Bayesian Reasoning and Machine Learning" by D. Barber Cambridge University Press (2012)
  http://www.cs.ucl.ac.uk/staff/d.barber/brml
- Material derived from slides for the book "Machine Learning" by T. Mitchell McGraw-Hill (1997)
  http://www- 2.cs.cmu.edu/~tom/mlbook.html
- Material derived from slides for the course "Machine Learning" by A. Ng, Stanford University, USA (2015)
- Material derived from slides for the course "Machine Learning" by A. Srinivasan BITS Pilani, Goa, India (2016)