Unbalanced Datasets

- For calculating accuracy we have implicitly assumed that there are the same number of positive and negative examples in the dataset (which is known as a balanced dataset).

balanced accuracy (BA)

$$\mathrm{BA} = \frac{TPR + TNR}{2}$$

- It doesn't hold good always

Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Example - MCC

- Given a sample of 13 pictures, 8 of cats and 5 of dogs, where cats belong to class 1 and dogs belong to class 0,
- Actual = [1,1,1,1,1,1,1,1,0,0,0,0,0],
- Prediction = [0,0,0,1,1,1,1,1,0,0,0,1,1]
- Classifier makes 8 accurate predictions and misses 5: 3 cats wrongly predicted as dogs (first 3 predictions) and 2 dogs wrongly predicted as cats (last 2 predictions).

		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

Example - MCC

- MCC = [(5*3) - (2*3)]/ SQRT[(5+2)*(5+3)*(3+2)*(3+3)] = 9/SQRT[1680] = 0.219

- Bayes Rule

$$P(C_i|X_j) = \frac{P(X_j|C_i)P(C_i)}{P(X_j)}$$

Example – Multiclass Bayesian classification

- What to do in the evening based on whether you have an assignment deadline and what is happening.

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

New Test Data: Suppose that you have deadlines looming, but none of them are particularly urgent, that there is no party on, and that you are currently lazy.

- P(Party) × P(Near | Party) × P(No Party | Party) × P(Lazy | Party)
- P(Study) × P(Near | Study) × P(No Party | Study) × P(Lazy | Study)
- P(Pub) × P(Near | Pub) × P(No Party | Pub) × P(Lazy | Pub)
- P(TV) × P(Near | TV) × P(No Party | TV) × P(Lazy | TV)

Using the data above these evaluate to:

$$
\begin{aligned}
P(\text{Party}|\text{near (not urgent) deadline, no party, lazy}) &= \frac{5}{10} \times \frac{2}{5} \times \frac{0}{5} \times \frac{3}{5} \\
&= 0 \\
P(\text{Study}|\text{near (not urgent) deadline, no party, lazy}) &= \frac{3}{10} \times \frac{1}{3} \times \frac{3}{3} \times \frac{1}{3} \\
&= \frac{1}{30} \\
P(\text{Pub}|\text{near (not urgent) deadline, no party, lazy}) &= \frac{1}{10} \times \frac{0}{1} \times \frac{1}{1} \times \frac{1}{1} \\
&= 0 \\
P(\text{TV}|\text{near (not urgent) deadline, no party, lazy}) &= \frac{1}{10} \times \frac{1}{1} \times \frac{1}{1} \times \frac{1}{1} \\
&= \frac{1}{10}
\end{aligned}
$$

So based on this you will be watching TV tonight.

Basic statistics

- Averages – mean, median, mode
- Variance
- The variance of the set of numbers is a measure of how spread out the values are. It is computed as the sum of the squared distances between each element in the set and the expected value of the set (the mean, μ)

$$\mathrm{var}(\{\mathbf{x}_i\}) = \sigma^2(\{\mathbf{x}_i\}) = E((\{\mathbf{x}_i\} - \boldsymbol{\mu})^2) = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^2$$

- Standard deviation
- https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc541.htm

Basic statistics

- Covariance
- If two variables are independent, then the covariance is 0 (the variables are then known as uncorrelated), while if they both increase and decrease at the same time, then the covariance is positive, and if one goes up while the other goes down, then the covariance is negative.

$$\operatorname{cov}(\{\mathbf{x}_i\}, \{\mathbf{y}_i\}) = E(\{\mathbf{x}_i\} - \boldsymbol{\mu})E(\{\mathbf{y}_i\} - \boldsymbol{\nu})$$

Basic statistics

Covariance matrix

The covariance can be used to look at the correlation between all pairs of variables within a set of data. We need to compute the covariance of each pair, and these are then put together into what is imaginatively known as the covariance matrix.

$$\Sigma = \begin{pmatrix} E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \\ E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \\ \dots & \dots & \dots & \dots \\ E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \end{pmatrix}$$

Example

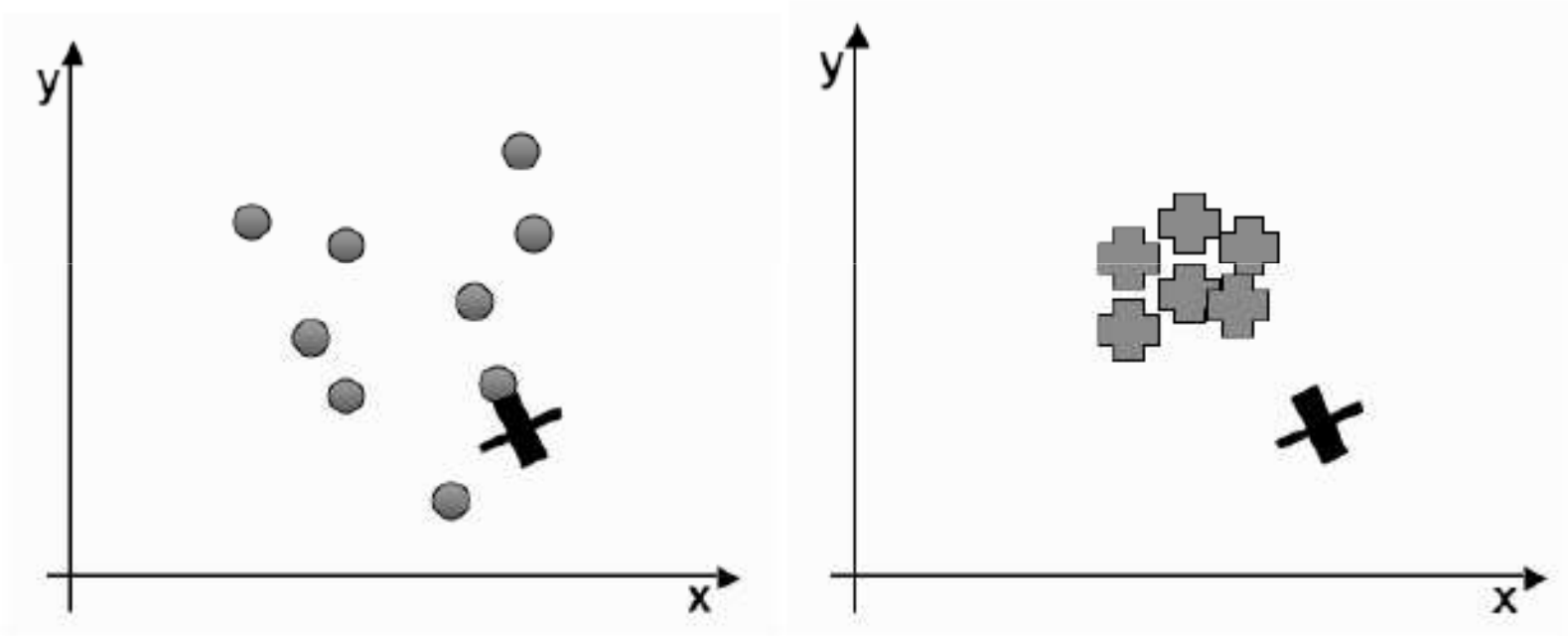


FIGURE 2.13 Two different datasets and a test point.

Gaussian Distribution

- One dimensional / multi dimential

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

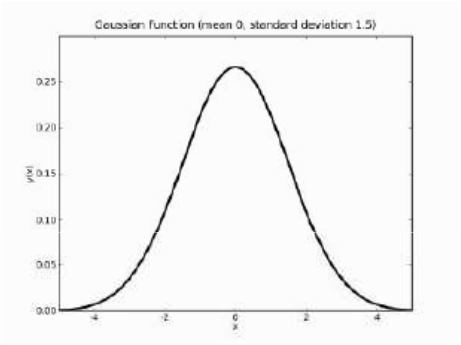


FIGURE 2.14 Plot of the one-dimensional Gaussian curve.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

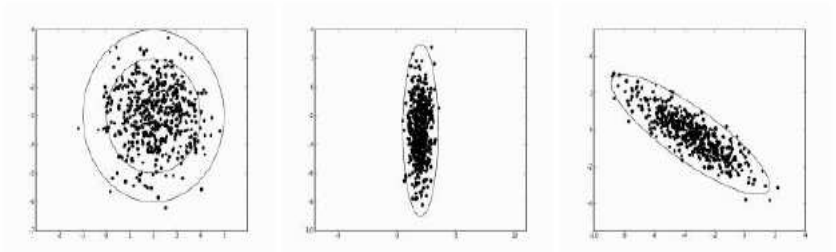
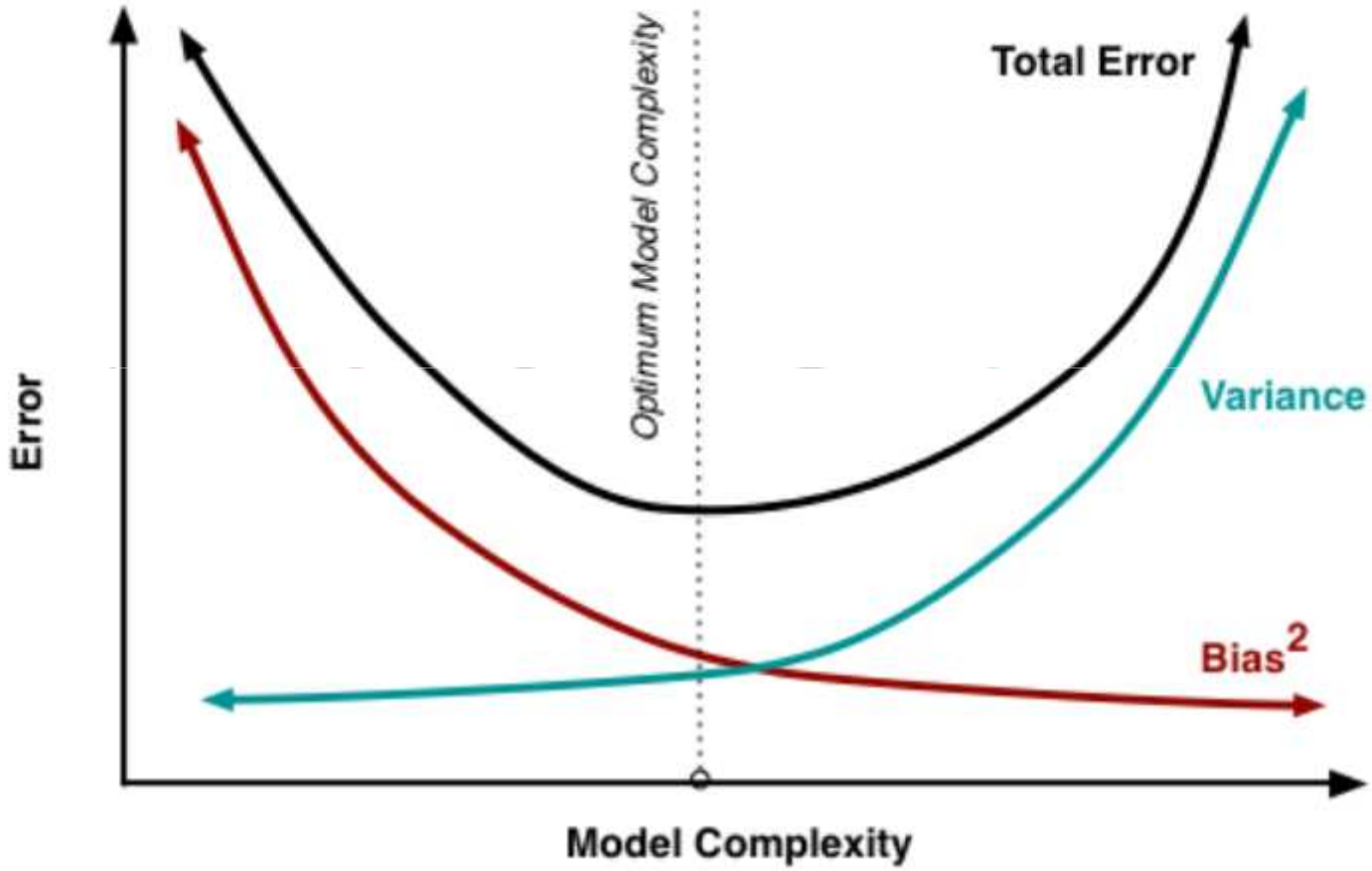


FIGURE 2.15 The two-dimensional Gaussian when (*left*) the covariance matrix is the identity, (*centre*) the covariance matrix has elements on the leading diagonal only, and (*right*) the general case.

Bias Variance Tradeoff

- Train ML – make choice about the model, fit parameters
- More degrees of freedom the algorithm has, the more complicated the model that can be fitted - overfitting, and requiring more training data, need for validation data
- There is another way to understand this idea that more complex models do not necessarily result in better results. Some people call it the **bias-variance dilemma** rather than a **tradeoff**.
- A model can be bad for **two different reasons**.
  - Either not accurate and doesn't match the data well -> bias
  - or it is not very precise and there is a lot of variation in the results -> statistical variance.

- Eg: Consider the difference between a straight line fit to some data and a high degree polynomial, which can go precisely though the datapoints.
- The straight line has no variance at all, but high bias since it is a bad fit to the data in general.
- The spline can fit the training data to arbitrary accuracy, but the variance will increase.
- Note that the variance probably increases by rather less than the bias decreases, since we expect that the
- spline will give a better fit.
- Some models are definitely better than others, but choosing **the complexity of the model is important** for getting good results.
- The variance - **x** changes depending on the particular training set that was used, while the bias - average error
- Expectation of sum of squares error = noise^2 + variance $+\text{bias}^2$



Bias and variance contributing to total error.