Probabilistic Graphical Models in Bioinformatics

Tutorial 3: Introduction to overfitting, Beta & Dirichlet distributions

Fahimeh Palizban
Department of Bioinformatics, IBB, University of Tehran

20 Esfand ,1397
(11 March ,2019)

University of TEHRAN

Contents

- What is overfitting?
- Beta distribution
- Dirichlet distribution
- Derivations

Overfitting

Definition of Overfitting

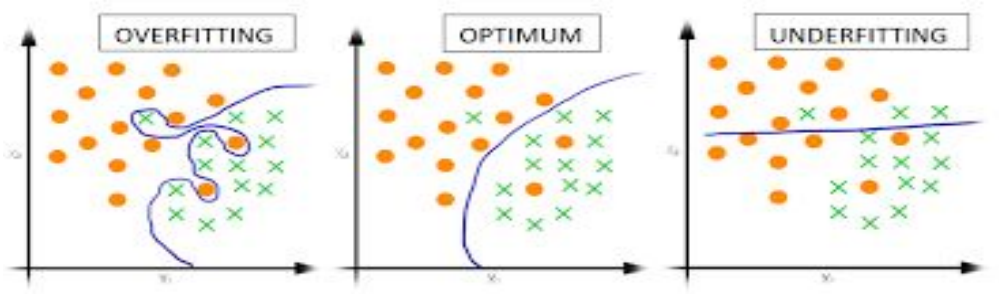
- Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points.
- An important consideration in learning the target function from the training data is how well the model generalizes to new data. Generalization is important because the data we collect is only a sample, it is incomplete and noisy.
- Generalization refers to how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning.

Statistical Fit

- In statistics, a fit refers to how well you approximate a target function.
- Statistics often describe the goodness of fit which refers to measures used to estimate how well the approximation of the function matches the target function.
- If we knew the form of the target function, we would use it directly to make predictions, rather than trying to learn an approximation from samples of noisy training data.

Overfitting in Machine Learning

- Overfitting refers to a model that models the training data too well.
- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function.

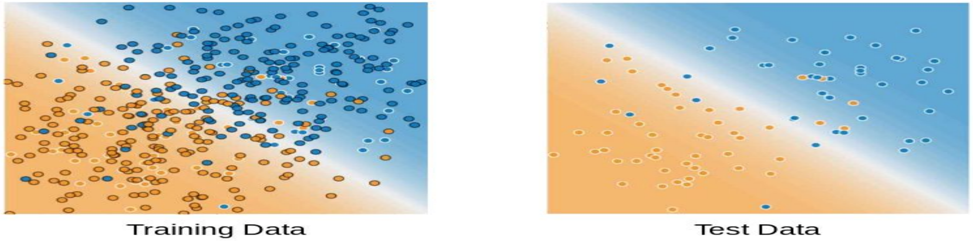


Split data to train & test

- **training set**—a subset to train a model.
- **test set**—a subset to test the trained model.

Make sure that your test set meets the following two conditions:

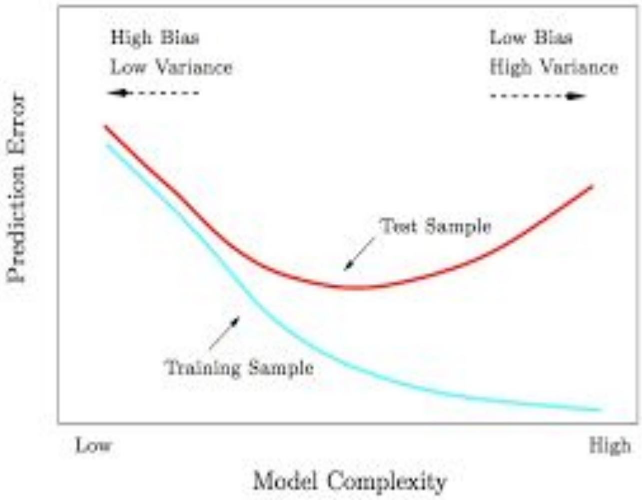
- Is large enough to yield statistically meaningful results.
- Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.



Tackling Overfitting

You must test your model on unseen data to counter overfitting.

A split of data 66%/34% for training to test datasets is a good start.



How To Address Overfitting

There are two important techniques that you can use when evaluating machine learning algorithms to limit overfitting:

Use a resampling technique to estimate model accuracy.

Hold back a validation dataset.

Feature selection

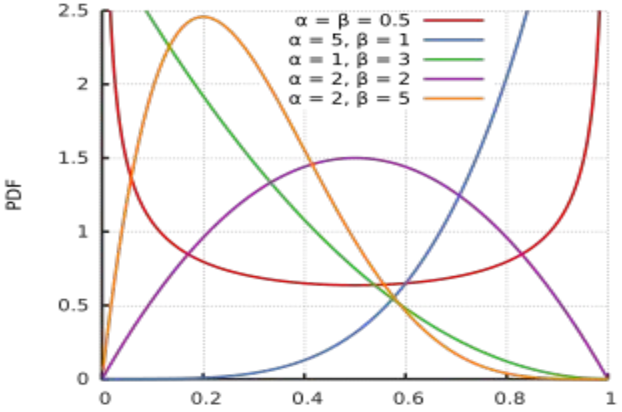
The most popular resampling technique is k-fold cross validation.

Beta distribution

Beta distribution

- A Beta distribution is used to model things that have a limited range, like 0 to 1.
- the beta distribution is conjugate prior to the Bernoulli distribution.

In short, the beta distribution can be understood as representing a probability distribution *of probabilities*



Beta Distribution

Notation	Beta(α, β)
Parameters	$\alpha > 0$ shape (real) $\beta > 0$ shape (real)
Support	$x \in [0, 1]$ or $x \in (0, 1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$ where $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
CDF	$I_x(\alpha, \beta)$ (the regularised incomplete beta function)
Mean	$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}$ $\mathrm{E}[\ln X] = \psi(\alpha) - \psi(\alpha+\beta)$ $\mathrm{E}[X \ln X] = \frac{\alpha}{\alpha+\beta}\left[\psi(\alpha+1) - \psi(\alpha+\beta+1)\right]$

Conjugate Prior

A conjugate prior is a probability distribution that, when multiplied by the likelihood and divided by the normalizing constant, yields a posterior probability distribution that is in the same family of distributions as the prior.

In other words, in the formula:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

The prior $p(\theta)$ is conjugate to the posterior $p(\theta|x)$ if both are in the same family of distributions.

For example, the normal distribution is conjugate to itself, because if the likelihood and prior are normal, then so is the posterior.

Dirichlet distribution

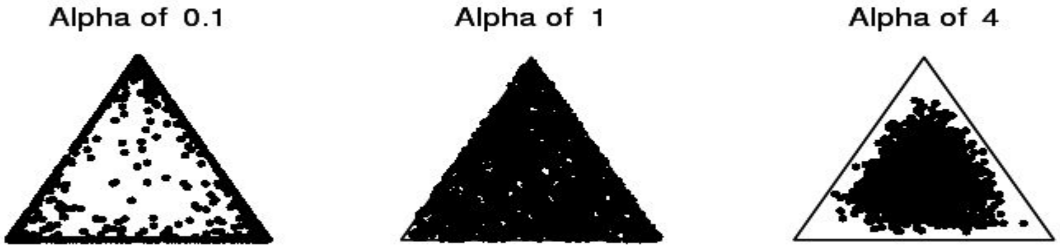
Dirichlet Distribution(the underlying intuition)

The Dirichlet distribution Dir(**α**) is a family of continuous multivariate probability distributions parameterized by a vector **α** of positive reals

It is a multivariate generalisation of the Beta distribution

Dirichlet distributions are commonly used as prior distributions in Bayesian statistics.

it is the *conjugate prior* to a number of important probability distributions: the categorical distribution and the multinomial distribution. Using it as a prior makes the maths a lot easier.



Parameters	$K \geq 2$ number of categories (integer) $\alpha_1, \ldots, \alpha_K$ concentration parameters, where $\alpha_i > 0$
Support	$x_1, \ldots, x_K$ where $x_i \in (0,1)$ and $\sum_{i=1}^{K} x_i = 1$
PDF	$\frac{1}{\mathrm{B}(\boldsymbol\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$ where $\mathrm{B}(\boldsymbol\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$ where $\boldsymbol\alpha = (\alpha_1, \ldots, \alpha_K)$
Mean	$\mathrm{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$ $\mathrm{E}[\ln X_i] = \psi(\alpha_i) - \psi(\sum_k \alpha_k)$ (see digamma function)
Mode	$x_i = \frac{\alpha_i - 1}{\sum_{k=1}^{K} \alpha_k - K}, \quad \alpha_i > 1.$
Variance	$\mathrm{Var}[X_i] = \frac{\tilde\alpha_i (1 - \tilde\alpha_i)}{\bar\alpha + 1},$ where $\tilde\alpha_i = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_i}$ and $\bar\alpha = \sum_{i=1}^{K} \alpha_i$ $\mathrm{Cov}[X_i, X_j] = \frac{-\alpha_i \alpha_j}{\bar\alpha + 1} \quad (i \neq j)$
Entropy	$H(X) = \log \mathrm{B}(\alpha) + (\alpha_0 - K)\psi(\alpha_0) - \sum_{j=1}^{K} (\alpha_j - 1)\psi(\alpha_j)$ with α_0 defined as for variance, above.

University of TEHRAN

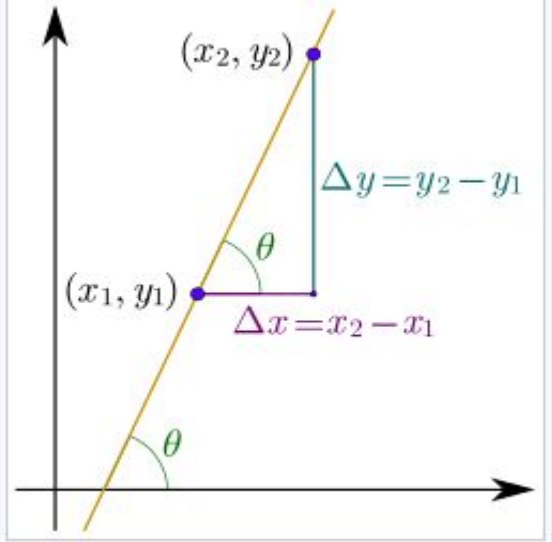
Derivation

Differentiation

Differentiation is the action of computing a derivative

The **derivative** of a function of a real variable measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value)

$$m = \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x},$$



Slope of a linear function: $m = \dfrac{\Delta y}{\Delta x}$

Basic Formulas of Derivatives

Common Functions	Function	Derivative
Constant	c	0
Line	x	1
	ax	a
Square	x^2	2x
Square Root	$\sqrt{x}$	$(½)x^{-½}$
Exponential	e^x	e^x
	a^x	$\ln(a)\ a^x$
Logarithms	ln(x)	1/x
	$\log_a(x)$	1 / (x ln(a))
Trigonometry (x is in radians)	sin(x)	cos(x)
	cos(x)	−sin(x)
	tan(x)	$\sec^2(x)$
Inverse Trigonometry	$\sin^{-1}(x)$	$1/\sqrt{(1-x^2)}$
	$\cos^{-1}(x)$	$-1/\sqrt{(1-x^2)}$
	$\tan^{-1}(x)$	$1/(1+x^2)$

University of TEHRAN

Thanks.