

Parameterized models

Curve fitting

(Big note of warning: for some reason PRML likes to call y the prediction and t the target, rather than calling y the target. I'm going to try to stick to this notation to be consistent with the book, but IMO it has it wrong.)

We can try to fit a M -degree polynomial:

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

to a bunch of data. Here, the $\mathbf{w} = \langle w_1, \dots, w_M \rangle$ term are the **parameters** of the model.

In order to fit the data, we introduce an error function, for example squared error. If we have data points $(x_1, t_1), (x_2, t_2), \dots, (x_N, t_N)$, then the error is:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$

We now try to minimize this as a function of \mathbf{w} (enter: calculus).

Different choices of M lead to different **models**, which can lead to:

- **Underfitting.** If M is too small, the model is too simple and we can't do a good job explaining the data.
- **Overfitting.** If M is too large, the model is too complex, and we can't do a good job generalizing.

These two things lead, in turn, to:

- High **training error.** We have a model that doesn't even do well on the training data!
- High **generalization error.** We have a model that doesn't generalize well to new points!

We notice that overfitting is *often* (though not always) discoverable by looking at the values of the weights: they tend to get huge! We can try to attenuate this by **regularizing** the model, for instance by penalizing large weights. We change our error function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Here, $\|\mathbf{w}\|^2 = \sum_j w_j^2$ is squared Euclidean norm. (Note that other regularizers are possible, and we'll talk about them later in the class.) Now, we can walk the line between underfitting and overfitting by tuning λ . λ is a **hyperparameter** of the model. We might tune it by **cross-validation** or by using **held-out data**.

Curse of Dimensionality

Big picture message: we're going to live in high dimensions for the rest of the semester and this gives us tons of parameters that we have to deal with.

The other parts (about how our intuitions fail in high dimensions) are particularly interesting.

The canonical ones are:

- Almost all of the volume of a sphere is concentrated in its shell for high dimensional spheres. (In the book.)
- Almost all of the mass of a high dimensional Gaussian distribution is concentrated in a thin shell. (In the book.)
- Spheres in high dimensions don't even look round! They look like porcupines! (See <http://mark.reid.name/iem/warning-high-dimensions.html>.)
- The ratio of the distance from the centre of the hypercube to one of the corners, divided by the perpendicular distance to one of the sides, is \sqrt{D} , which therefore goes to ∞ as D goes to ∞ . From these results we see that, in a space of high dimensionality, most of the volume of a cube is concentrated in the large number of corners, which themselves become very long spikes. (This is an exercise in PRML.)

Take away message: high dimensions are weird.