ML Notes

1 Introduction

Goal: Take a series of (labelled) examples, represented by their features and adapt the parameters of a model, s.t. A function $y(x)$ takes in a new data point represented in the same way as the training data points and outputs a vector that categorises it.

1.1 Types of Learning

1.1.1 Supervised

Learn from labelled data: classification (assigning input vector to discrete categories), regression (assigning to continuous variables). Need pairs of training data points with labels to train on.

1.1.2 Unsupervised

Learn from unlabelled data: clustering (discover groups of similar examples), density estimation (discover how data is distributed in input space)

1.1.3 Reinforcement

Find which actions should be taken to maximise a reward. Trial and error process, where there is a sequence of states and actions, where a current action affects immediate reward at a state, but also affects rewards at later timesteps. Intuitively, a network can learn by interacting with itself or other agents, until the optimal set of moves is achieved.

One problem is tradeoff between exploration (trying out possible actions) and exploitation (prioritise actions that yield high reward).): e.g. credit assignment problem

2 Examples

2.1 Polynomial Curve Fitting

Goal: given training set data, where each data point is mapped to a target variable (and the data points come from some function), our goal is to implicitly try to discover the underlying function. This is difficult, because we want to generalise from a finite data set.

$y(x, w) = w_o + w_1 x + w_2 x^2 + .. + w_m x^m = \sum_{j=0}^{m} w_j x^j$

Although the polynomial is non-linear in x, it is a linear function of the coefficients. The value of the coefficients is determined by trying to fit the polynomial to the training data by minimising an error function. For example, one error function is the sum of squares of errors of predictions vs. the gold values. We therefore want to minimise:

$E(w) = \frac{1}{2} \sum_{i=1}^{n} \{y(x_i, w) - t_i\}^2$

We want to choose coefficients w to minimise $E(w)$. Since the error is quadratic in w, it's derivate is linear, so there exists a unique solution. We also have to choose the order of the polynomial in the first place and this is called model comparison or model selection. We sometimes use the RMS error $E_{RMS} = \sqrt{2E(w)/N}$ so that we can compare different sizes of datasets. The coefficients may sometimes be too finely tuned to the training data, so we use a penalty term or regularisation to control the over fitting phenomenon.

$E(w) = \frac{1}{2} \sum_{i=1}^{n} \{y(x_i, w) - t_i\}^2 + \frac{\lambda}{2} \|w\|^2$

Here λ governs the relative importance of the regularisation penalty against the sum of squares error.

2.2 Gaussian Distribution

This is a probability distribution governed by two factors, the mean μ and variance σ^2. It is defined by:

$\aleph\ (x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x - \mu^2))$

THe Gaussian is normalised so that:

$$\int_{-\infty}^{\infty} \aleph\ (x|\mu, \sigma^2) = 1$$

Now given a set of observations $\mathbf{x} = (x_1, x_2, ..., x_N)^T$, such that observations are drawn independently from the same distribution and are therefore *i.i.d.*, the probability of the data et, given μ and σ^2 is:

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \aleph(x_n|\mu, \sigma^2).$$

This is the likelihood function for the Gaussian. We can determine values for the unkown parameters μ and σ^2 to maximise the likelihood function. Intuitievly, we should note that maximising the probability of the parameters given the data and the probability of the data given the parameters are closesly related. Instead of working with the likelihood, we work with the log likelihood, and maximising this.

Maximising wrt μ, we get:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1} N x_n.$$

Maximising wrt σ^2, we get:

$$\sigma^2{}_{ML} = \frac{1}{N}\sum_{n=1} N(x_n = \mu_{ML})^2.$$

The limitations of the maximum likelihood approach are that is systematically underestimates the variance of the distribution i.e. causes a bias. If we look at the expectations of μ_{ML} and $\sigma^2{}_{ML}$, we see that:

$$E[\mu_{ML}] = \mu$$

$$E[\sigma^2{}_{ML}] = \frac{N-1}{N}\sigma^2.$$

Therefore on average, the maximum likelihood estimate will obtain the correct mean but will udnerestimate the true variance by a factor of $\frac{N_1}{N}$