Task 1.2. Supervised Learning: Bayesian Linear Regressor

Garoe Dorta-Perez
CM50246: Machine Learning and AI

January 16, 2015

1 Introduction

Simple linear regressors are overconfident in their predictions. Bayesian linear regressors are an extension of this model used to solve this particular issue.

2 The problem

In this model we want to compute a posterior distribution given a set of training samples, as shown in Equation 1, where $\mathbf{w}$ is a one dimensional array with the world state, $\mathbf{X}$ is a matrix with the data points, ϕ are the parameters of a linear function of the data.

$$Pr(\phi|\mathbf{X},\mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X},\phi)Pr(\phi)}{Pr(\mathbf{w}|\mathbf{X})}. \tag{1}$$

The prior in Equation 1 is a normal distribution with 0 mean and spherical covariance. The likelihood is a multivariate normal distribution, as shown in Equation 2. Where σ^2 is the covariance, $\mathbf{I}$ is the identity matrix and $\boldsymbol{\theta} = \{\phi, \sigma^2\}$.

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\theta}) = Norm_{\mathbf{w}}\left[\mathbf{X}^T\phi, \sigma^2\mathbf{I}\right]. \tag{2}$$

The posterior distribution is shown in Equation 3.

$$\begin{gathered} Pr(\phi|\mathbf{X},\mathbf{w}) = Norm_{\phi}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right], \\ \mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}. \end{gathered} \tag{3}$$

The probability distribution of a new world state w^* over new test data x^* is a Normal distribution as shown in Equation 4.

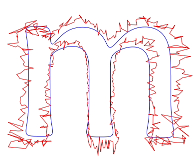
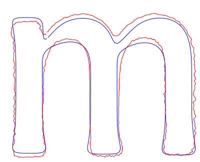
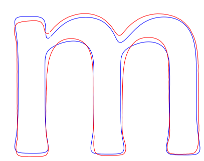
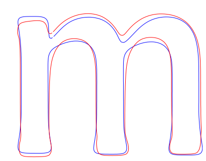
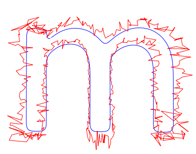
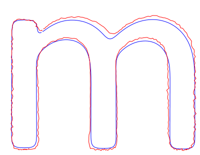
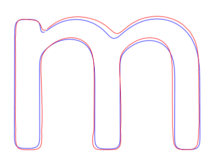
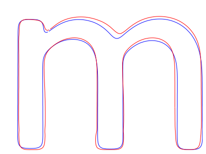
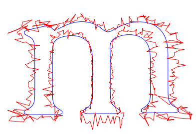
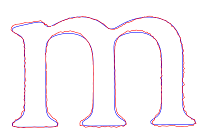
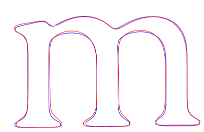
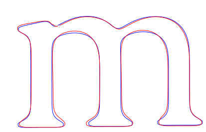
$$
\begin{aligned}
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \phi) Pr(\phi|\mathbf{X}, \mathbf{w}) d\phi, \\
&= Norm_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2\right].
\end{aligned} \tag{4}
$$

Since our world states $\mathbf{w_i}$ are multivariate, a separate regressor is used for each dimension in $\mathbf{w}$.

3 Results

In Table 1 we show results for a font data set. The Bayesian linear regressors is trained with pictures of n in different fonts and their corresponding m in the same font. For testing, n with new fonts are used.

Table 1: Comparing the reconstruction of several m with different fonts and variances. In blue is the ground truth and in red is the reconstructed character.

	$\sigma^2 = 10$	$\sigma^2 = 1$	$\sigma^2 = 0.01$	$\sigma^2 = 0.001$
Font1				
Font2				
Font3				

As shown in Table 1, smaller variances σ^2 produce better results. Therefore, our prior model of a Normal distribution is hardly useful for this dataset, and we are better off relying only on the data. The quantity of samples in the dataset is relatively small, this is even more evident if we compare it against the data dimensionality, leading to poor generalization, for instance there are not samples of the top arch in Font2 in the training data.