Alin Tomescu
6.867 Machine learning | Week 4, Thursday, September 26th, 2013| Lecture 7

Lecture 7: Statistical modeling

Model is always a set.

$$y = \theta\phi(x) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

Assuming $N(0, \sigma^2)$ is fixed, we can compute $P(y \mid x, \theta)$:

$$P(y \mid x, \theta) = N(y; \theta\phi(x), \sigma^2)$$

- the y in $N(y; \theta\phi(x), \sigma^2)$ indicates y is the random variable
- $\theta\phi(x)$ is the mean
- σ^2 is the variance.

$$N(y; \theta\phi(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\theta\phi(x))^2}$$

$$y^{(1)} = \phi\left(x^{(1)}\right) + \varepsilon_1$$

$$\vdots$$

$$y^{(n)} = \phi\left(x^{(n)}\right) + \varepsilon_n$$

Maximum likelihood

Find θ that maximizes the likelihood function $L(\theta; S_n)$:

$$P(y \mid x, \theta) = L(\theta; S_n) = \prod_{i=1}^{n} P\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

$$\max L(\theta; S_n) = \max \log L(\theta; S_n) = l(\theta; S_n) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{1}{2}\left(y^{(i)} - \phi\left(x^{(i)}\right)\right)^2 + \text{ constant}$$

Maximum a posteriori estimation

Find θ that maximizes $L(\theta; S_n)P(\theta)$, where $P(\theta)$ is the prior.

A typical prior is $P(\theta) = N(\theta; 0, v^2 I) = \frac{1}{(2\pi v^2)^{\frac{d}{2}}} e^{-\frac{1}{2v^2}\|\theta\|^2}$, where d is the dimension of θ (see fig. 2) (this is a multivariate Gaussian)

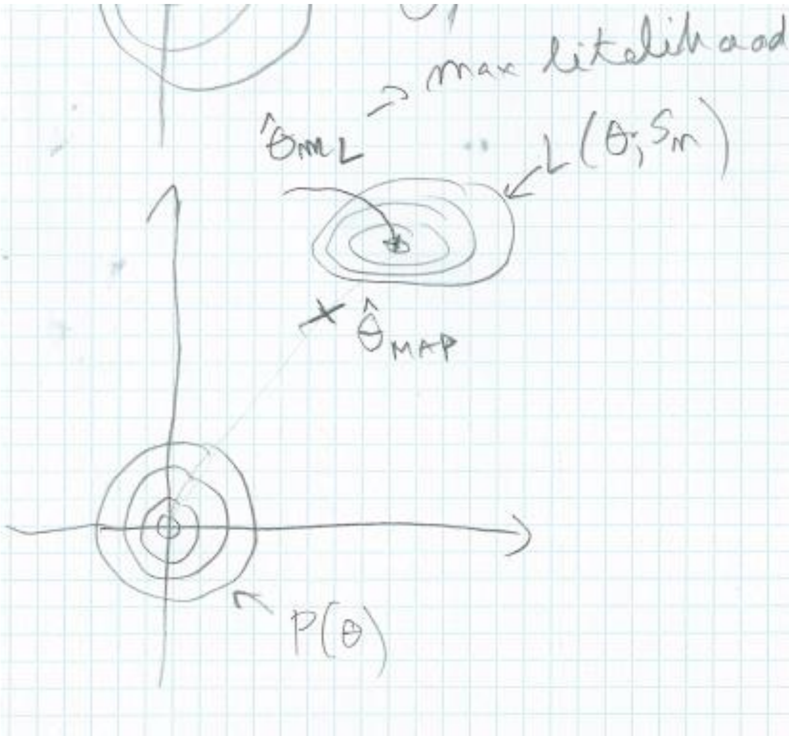
$$l(\theta; s_n) = \log L(\theta, S_n) + \log P(\theta) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{1}{2}\left(y^{(i)} - \phi\left(x^{(i)}\right)\right)^2 - \frac{1}{2v^2}\|\theta\|^2 + \text{constant}$$

$$-\frac{1}{\sigma^2}\left[\sum_{i=1}^{n} \frac{1}{2}\left(y^{(i)} - \phi\left(x^{(i)}\right)\right)^2 + \frac{1}{2}\frac{\sigma^2}{v^2}\|\theta\|^2\right] + \text{constant}$$

Alin Tomescu

6.867 Machine learning | Week 4, Thursday, September 26th, 2013| Lecture 7

Regularization parameter is: $\lambda = \frac{\sigma^2}{\nu^2}$



Bayesian estimator

$$P(\theta \mid S_n) = \frac{1}{Z(S_n)} L(\theta; S_n) P(\theta)$$

(what is $Z(S_n)$?) Z is just a normalization constant.

$$Z(S_n) = \int_{\mathbb{R}^d} L(\theta; S_n) P(\theta) d\theta = \int_{\mathbb{R}^d} \prod_{i=1}^{n} P\left(y^{(i)} \mid x^{(i)}, \theta\right) d\theta = P\left(y^{(1)}, y^{(2)}, \dots, y^{(n)} \mid x^{(1)}, x^{(2)}, \dots, x^{(n)}, \mathcal{M}\right) = Z(S_n, \mathcal{M})$$

This is **marginal likelihood**.

$$P\left(y^{(1)}, y^{(2)}, \dots, y^{(n)} \mid x^{(1)}, x^{(2)}, \dots, x^{(n)}, \mathcal{M}\right)$$

No longer depends on θ, but it depends on the model. What is the model here?

$$\mathcal{M} = \{N(y; \theta\phi(x), \sigma^2), P(\theta), \theta \in \mathbb{R}^d\}$$

Model selection:

$$\mathcal{M}_1: \phi(x) = x$$

$$\mathcal{M}_2: \phi(x) = \begin{bmatrix} x^2 \\ x \end{bmatrix}$$

$$\vdots$$

For all of these I can evaluate: $Z(S_n; \mathcal{M}_1), Z(S_n; \mathcal{M}_2)$

We will select $\mathcal{M}$ that maximizes this marginal likelihood $Z(S_n, \mathcal{M}_i)$

Bayesian information criterion

$$\log Z(S_n; \mathcal{M}) \cong \log L(\widehat{\theta_{ML}}; S_n) - \frac{d}{2}\log n$$

Penalty: $\frac{d}{2}\log n$, where d is the # of parameters.

Asymptotic expansion (leading order term $\log n$).

When n is large and d is smaller, this is good.

We no longer depend on the prior. What happened to the prior? As n increases, the prior no longer dominates.

Bayesian prediction

Data $S_n \Rightarrow P(\theta; S_n)$

I have a new x, what is $P(y \mid x, S_n) = ?$

If we were doing ML, then we'd use $P(y \mid x, \widehat{\theta_{ML}})$

$$P(y \mid x, S_n) = \int_{\mathbb{R}^d} P(y \mid x; \theta)\, P(\theta \mid S_n) d\theta = \int_{\mathbb{R}^d} P(y \mid x; \theta) \frac{\left[\prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}, \theta)\right] P(\theta)}{Z(S_n)} d\theta$$
$$= \frac{1}{Z(S_n)} P(y^{(1)}, y^{(2)}, \dots, y^{(n)} \mid x^{(1)}, x^{(2)}, \dots, x^{(n)}, \mathcal{M})$$

$$\theta \sim P(\theta) = N(\theta; 0, \nu^2 I)$$

$$\vdots$$

$$y^i = \theta\phi(x^{(i)}) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

$$\vdots$$

Gaussian process

A Gaussian process is a collection of random variables $\{y_x, x \in \mathcal{X}\}$ if $\forall n$ $\{X^1, \dots, X^{(n)}\}$ and $\{y^{(i)} = y_{x^{(i)}}\}$ are jointly Gaussian.