CMPUT 367

Roderick Lan

Contents

1 Lecture 3 - Jan 18 **2**
- 1.1 Maximum a Posteriori (MAP) 2
- 1.2 Assignment 3
- 1.3 Binary Classification 3
- 1.4 Multiclass Classification 4

1 Lecture 3 - Jan 18

$$t^{(m)} \sim \mathcal{N}(w^\top x^{(m)}, \sigma^2)$$

$\mathcal{D}_{\text{train}} = \{(x^{(m)}, t^{(m)})\}_{m=1}^{M}$ (training set w/ m samples)

$$\hat{w}_{MLE} = \arg\max_w p(D; w)$$

w is not a random variable

$$\begin{aligned}
&\arg\max_w p(t^{(1)}, \ldots, t^{(m)} | x^{(1)}, \ldots, x^{(m)}; w) \\
=& \arg\max_w \prod_{m=1}^{M} p(t^{(m)} | x^{(1)}, \ldots, x^{(m)}; w) \\
=& \arg\max_w \prod_{m=1}^{M} p(t^{(m)} | x^{(m)}; w)
\end{aligned}$$

$t^{(m)}$ only related to $x^{(m)}$; can drop other $x^{(i)}$

$$\begin{aligned}
=& \arg\max_w \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{t^{(m)} - w^\top x^{(m)}}{\sigma}\right)^2\right\} \\
=& \arg\max_w \log \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{t^{(m)} - w^\top x^{(m)}}{\sigma}\right)^2\right\} \\
=& \arg\min_w \left(\frac{1}{2\sigma^2} \sum_{m=1}^{M} (t^{(m)} - w^\top x^{(m)})^2\right) \\
=& \arg\min_w \left(\frac{1}{2M} \sum_{m=1}^{M} (t^{(m)} - w^\top x^{(m)})^2\right)
\end{aligned}$$

Since σ^2 is a constant; irrelevant, can be replaced with anything

$$= \arg\min_w (MSE)$$

MSE $\iff$ MLE assuming $t^{(m)} \sim \mathcal{N}(w^\top x^{(m)}, \sigma^2)$

1.1 Maximum a Posteriori (MAP)

w = random variable

$$t^{(m)} \sim \mathcal{N}(w^\top x^{(m)}, \sigma^2)$$

$$w_i \sim \mathcal{N}(0, \sigma_{noise}^2) \text{ or } w_i \sim \text{Laplacian}(0, \lambda)$$

$$\hat{w}_{MAP} = \arg\max p(w|\mathcal{D})$$

$$
\begin{aligned}
\hat{w}_{MAP} &= \arg\max_w p(w|\mathcal{D}) \\
&= \arg\max_w \frac{p(\mathcal{D}|w)\cdot p(w)}{p(\mathcal{D})} \\
&= \arg\max_w p(\mathcal{D}|w)p(w) \\
&= \arg\max_w \log p(\mathcal{D}|w)p(w) \\
&= \arg\max_w [\log p(\mathcal{D}|w) + \log p(w)]
\end{aligned}
$$

$l1$-penalty if $w \sim \mathcal{N}$
$l2$-penalty if $w \sim$ Laplacian

Explain 1.1.1: Random Variable

Frequentist Interpretation - a RV is the outcome of a repeatable experiment
Bayesian Interpretation - anything unknown can be a RV: subjective belief

1.2 Assignment

MAP w/ Gaussian - L2-regularization (dense model)

$$
J(w) = \frac{1}{2M}\|Xw - t\|^2 + \lambda\|w\|_2^2
$$

MAP w/ Laplacian - L1-regularization (sparse model)

$$
J(w) = \frac{1}{2M}\|Xw - t\|^2 + \lambda\|w\|_1
$$

soft penalty equivalent to a hard constraint in convex optimization
large hypothesis class implies overfitting, regularization constrains hypothesis class.
MLE $\iff$ MSE
MAP $\iff$ MSE + reg

1.3 Binary Classification

$$
t^{(m)} \sim \{0,1\}
$$

$$
t^{(m)} \sim \text{Bernoulli}\left(\sigma(w^\top x^{(m)} + b)\right)
$$

$$
\sigma(z) = \frac{1}{1+e^{-z}}
$$

$$\mathcal{D}_{train} = \{(x^{(m)}, t^{(m)})\}_{m=1}^{M}$$

$$\begin{aligned}
p(\mathcal{D}; w, b) &= \prod_{m=1}^{M} p(t^{(m)}|x^{(m)}; w, b) \\
&= \arg\max_{w,b} \prod_{m=1}^{M} \sigma(w^\top x^{(m)} + b)^{(t^{(m)})}(1 - \sigma(w^\top x^{(m)} + b)) \\
&= \arg\max \ln \prod_{m=1}^{M} \sigma(w^\top x^{(m)} + b)^{(t^{(m)})}(1 - \sigma(w^\top x^{(m)} + b))^{1-t^{(m)}} \\
&= \arg\max \sum_{m=1}^{M} [(t^{(m)}) \ln \sigma(w^\top x^{(m)} + b) + (1 - t^{(m)}) \ln(1 - \sigma(w^\top x^{(m)} + b))] \\
&= \arg\max(\text{Cross Entropy Loss})
\end{aligned}$$

if $t^{(m)} = 1$, likelihood: $\sigma(w^\top x^{(m)} + b)$
elif, $t^{(m)} = 0$, likelihood: $1 - \sigma(w^\top x^{(m)} + b)$

1.4 Multiclass Classification

Probability Assumption (form of dist.)
Parameter
Likelihood
Loss