Lecture 10: 2/20/17

Regularization

Goal: prefer "simpler" model parameters

L0, L1, L2 Regularization:

Suppose: $\ell(\vec{\theta})$ is a likelihood fn.

Examples

$\ell(\vec{\theta}) \triangleq \log \prod_{i=1}^{N} p(x^{(i)}, y^{(i)} | \vec{\theta})$ ← NB

$\ell(\vec{\theta}) \triangleq \log \prod_{i=1}^{N} p(y^{(i)} | x^{(i)}, \vec{\theta})$ ← Log. Reg.

$\ell(\vec{\theta}) \triangleq \log \prod_{i=1}^{N} g(y^{(i)} | x^{(i)}, \vec{\theta})$ ← Lin. Reg.

Define:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ J(\theta)$$

$$J(\theta) = -\ell(\theta) + \text{"model complexity"}$$

$$= \underbrace{-\ell(\theta)}_{\text{neg. log. likelihood}} + \lambda \underbrace{r(\theta)}_{\text{regularization}}$$

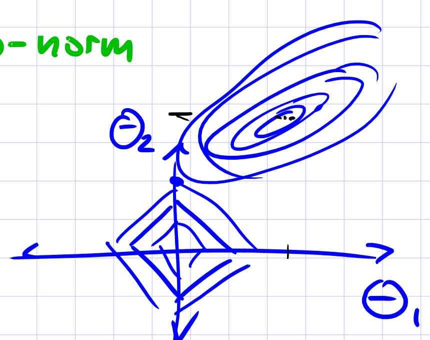
λ: tunable parameter chosen on validation

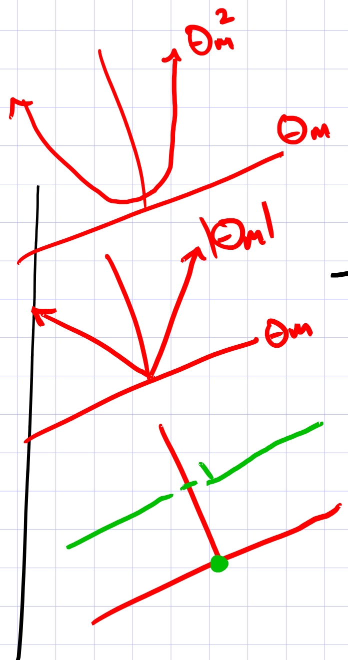
Key Idea: Define $r(\theta)$ s.t. we tradeoff between fitting the data and keeping the model simple.

Choose form of $r(\vec{\theta})$:

Usually $r(\vec{\theta}) \triangleq \|\vec{\theta}\|_q$ — typically "p" i.e. p-norm

$$= \left[\sum_{M=1}^{M} |\theta_m|^q\right]^{1/q}$$





q	$r(\theta)$	Preference for...	Notes
2	$(\|\theta\|_2)^2 = \sum \theta_m^2$	Small values.	[L2 reg., differentiable]
1	$\|\theta\|_1 = \sum \|\theta_m\|$	zero values.	[L1 reg., subdifferentiable]
0	$\|\theta\|_0 = \sum \mathbb{1}(\theta_m != 0)$	zero values	[L0 reg., no good computational solutions]

Example: Linear Regression

$r(\theta) = (\|\theta\|_2)^2 \implies$ L2 reg. aka. "Ridge Regression"

$r(\theta) = (\|\theta\|_1) \implies$ L1 reg. aka "LASSO"

$r(\theta) = (\|\theta\|_0) \implies$ L0 reg. aka "Subset Selection"

Prob. Interp. of Regularization

Punchline: L2 reg. is MAP estimation w/ Gaussian prior
L1 reg. is MAP estimation w/ Laplace prior.

$$\ell_{MAP}(\theta) = \log[p(D|\theta)\,p(\theta)]$$
$$= \log p(D|\theta) + \underline{\log p(\theta)}$$

Ex: Zero-mean Gaussian prior on $\vec{\theta}$

Story:

$$\theta_m \sim \text{Gaussian}\left(\mu = 0, \sigma^2 = \frac{1}{2\lambda}\right) \quad \forall m$$
$$D \sim p(D|\theta)$$

$$\ell_{MAP}(\theta) = \log p(D|\theta) + \log\left[\prod_{m=1}^{M} f_{Gaussian}\left(\theta_m \,\middle|\, \mu = 0, \sigma^2 = \frac{1}{2\lambda}\right)\right]$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \ell_{MAP}(\vec{\theta})$$
$$= \underset{\theta}{\operatorname{argmax}}\ \log p(D|\theta) + \log\left[\prod_{m=1}^{M} f_{Gaussian}\left(\theta_m \,\middle|\, \mu = 0, \sigma^2 = \frac{1}{2\lambda}\right)\right]$$
$$= \underset{\theta}{\operatorname{argmax}}\ \log p(D|\theta) + \sum_{m=1}^{M} \cancel{-\log\left(\sqrt{2\sigma^2\pi}\right)} - \frac{1}{2\sigma^2}(\theta_m)^2$$
$$= \underset{\theta}{\operatorname{argmax}}\ \log p(D|\theta) - \sum_{m=1}^{M} \frac{1}{2\sigma^2}(\theta_m)^2$$
$$= \underset{\theta}{\operatorname{argmax}}\ \log p(D|\theta) - \underbrace{\lambda \sum_{m=1}^{M} \theta_m^2}_{\text{L2 reg.}}$$

Ex: Zero-mean Laplace prior on $\vec{\theta}$

Story:

$$\theta_m \sim \text{Laplace}\left(\mu = 0,\ b = \frac{1}{\lambda}\right) \quad \forall m$$

$$D \sim p(D \mid \vec{\theta})$$

$\Rightarrow$ $\log p(\theta)$ is equiv. to $\lambda \|\theta\|_1$ penalty