Lecture 18, part I

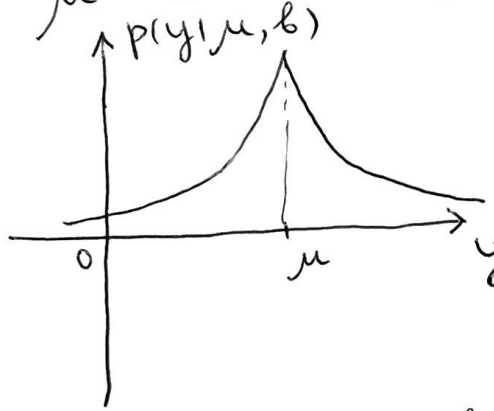
Robust linear regression

Fitting gaussian models is sensitive to outliers $\Rightarrow$ one can use heavy-tail distributions to achieve robustness to outliers.

(1.) Laplace likelihood

Laplace distribution: $p(y|\mu,b) = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$.

μ = location prm, b = scale prm. $[b>0]$



$\int_{-\infty}^{\infty} dy\, p(y|\mu,b) = 1$

$\begin{cases} \text{mean} = \text{mode} = \mu, \\ \text{var} = 2b^2 \end{cases}$

Needs techniques like projected grad for robust optimization.

(2.) Student's t-likelihood

Student's t-distribution:

$$p(y|\mu,\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{1}{\nu}\left(\underbrace{\frac{y-\mu}{\sigma}}_{\text{"}t}\right)^2\right)^{-\frac{\nu+1}{2}}.$$

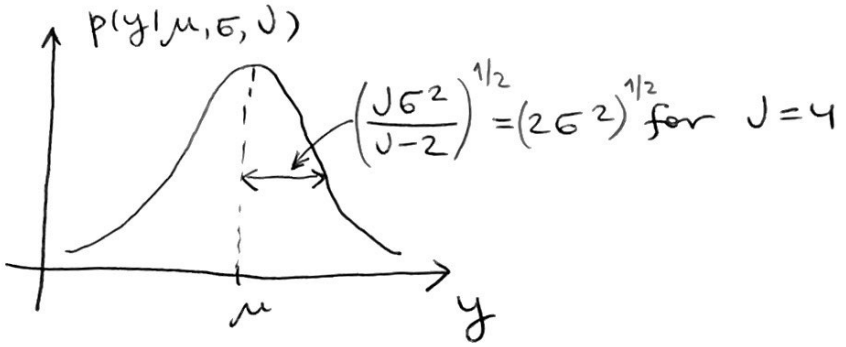
μ = mean prm, $\sigma > 0$ = scale prm, $\nu > 0$ = DoF

$$\begin{cases} \text{mean} = \mu, & \nu > 1 \\ \text{mode} = \mu \\ \text{var} = \dfrac{\nu\sigma^2}{\nu-2}, & \nu > 2 \end{cases} \Rightarrow \nu = 4 \text{ is commonly used}$$

Note that for $t^2 \gg \nu$,

$$p(t) \sim \left(t^2\right)^{-\frac{\nu+1}{2}} = t^{-(\nu+1)}$$

Thus, the tail decays much slower than the gaussian tail.



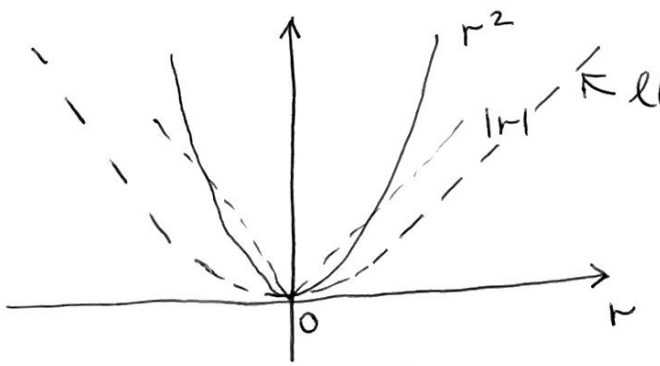
The model can be optimized using SGD or other standard techniques.

(3.) Huber loss

$$\ell(r) = \begin{cases} r^2/2, & |r| \le \delta \\ \delta|r| - \delta^2/2, & |r| > \delta \end{cases}$$

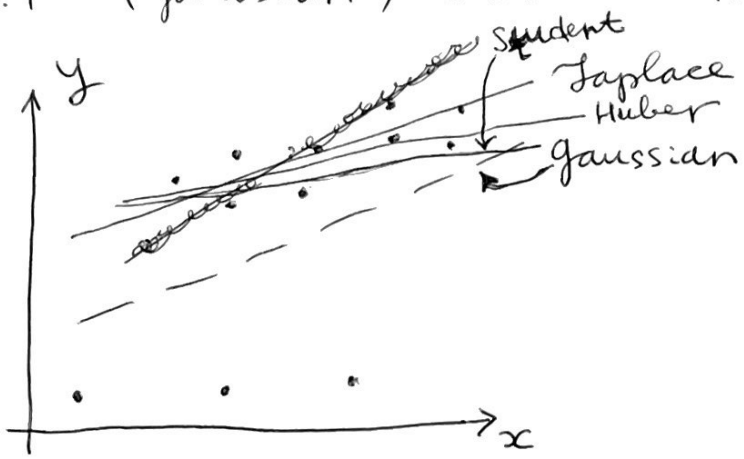
$r = w - a$ ← fitting weight; ↑ constraint, typically $= 0$,

(or) r = obs. − model $y - \hat{y}$



Typically, $\delta = \Theta(1)$, e.g. 1.5

Huber loss can be used instead of $|\dots|^2$ (gaussian) loss or $|\dots|$ (Laplace) loss.



Bayesian linear regression

Likelihood: $p(\mathcal{D}|\vec{w}, \sigma^2) = \prod_{n=1}^{N} p(y_n | \vec{w}\cdot\vec{x}, \sigma^2) =$

$= \mathcal{N}(\vec{y} | X\vec{w}, \sigma^2 \mathbb{I}_N)$. centered data

Prior: $p(\vec{w}) = \mathcal{N}(\vec{w} | \vec{0}, \tau^2 \mathbb{I}_D)$.

Then the posterior is given by

$p(\vec{w}|\mathcal{D}, \sigma^2) = \mathcal{N}(\vec{w} | \vec{w}', \Sigma')$, where

$$\begin{cases} \vec{w}' = \frac{1}{\sigma^2} \Sigma' X^T \vec{y}, \\ \Sigma' = \left(\frac{1}{\tau^2}\mathbb{I}_D + \frac{1}{\sigma^2} X^T X\right)^{-1} = \sigma^2 \left(\underbrace{\lambda}_{\frac{\sigma^2}{\tau^2}} \mathbb{I}_D + X^T X\right)^{-1}. \end{cases}$$

Then $\vec{w}' = (\lambda \mathbb{I}_D + X^T X)^{-1} X^T \vec{y}$,

same as ridge regression.

Finally, we can compute the predictive distribution:

$$p(y|\vec{x}, \mathcal{D}, \sigma^2) = \int d\vec{w}\, \mathcal{N}(y|\vec{w}\cdot\vec{x}, \sigma^2) \times$$

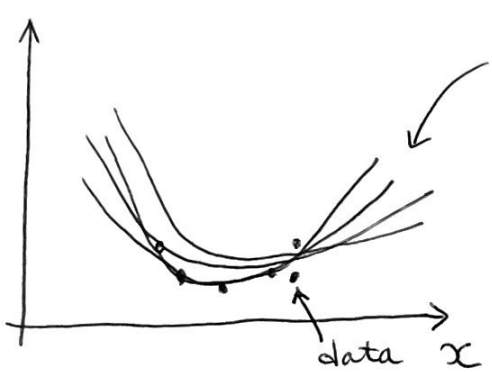
$$\times \mathcal{N}(\vec{w}|\vec{w}', \Sigma') =$$

$$= \mathcal{N}(y|\vec{w}'\cdot\vec{x}, \hat{\sigma}^2(\vec{x})), \quad \text{where}$$

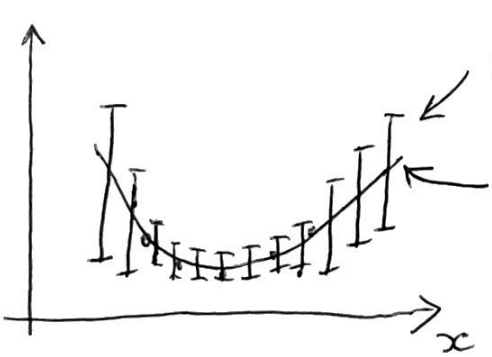
$$\hat{\sigma}^2(\vec{x}) = \sigma^2 + \vec{x}^T \Sigma' \vec{x}.$$

σ^2: var. in observation noise

$\vec{x}^T \Sigma' \vec{x}$: var. in posterior prms after seeing N training examples



$f(x) = wx^2$, w sampled from the posterior



$\pm 2\hat{\sigma}(x)$ error bars

mean of predictive distribution $= w'x^2$

Now, we switch to features:

$$\underset{N\times D}{\mathbf{X}} \Rightarrow \underset{N\times M}{\Phi}$$

data matrix — design matrix

and switch to Bishop's notation:

posterior $\Rightarrow$ $p(\vec{w}\,|\,\mathcal{D},\sigma^2)$ becomes

$$p(\vec{w}\,|\,\mathcal{D},\underset{\tau^{-2}}{\alpha},\underset{\sigma^{-2}}{\beta}) = \mathcal{N}(\vec{w}\,|\,\vec{m}_N, S_N), \text{ where}$$

$\mathcal{D} = \{\vec{t}, X\}$ ← instead of $\vec{y}$

$$\begin{cases} \vec{m}_N = \beta S_N \Phi^T \vec{t}, \\ S_N = (\alpha \mathbb{I} + \beta \Phi^T \Phi)^{-1} \end{cases}$$

Thus, we have relabeled:

$$\begin{cases} \vec{w}' \Rightarrow \vec{m}_N, \\ \Sigma' \Rightarrow S_N. \end{cases}$$

Finally, the predictive distribution is given by

$$p(t\,|\,\vec{x},\mathcal{D},\alpha,\beta) = \mathcal{N}(t\,|\,\vec{m}_N\cdot\underbrace{\vec{\varphi}(\vec{x})}_{\text{M-dim feature vector}}, \sigma_N^2(x)),$$

where

$$\sigma_N^2(\vec{x}) = \beta^{-1} + \vec{\varphi}^T(\vec{x}) S_N \vec{\varphi}(\vec{x}).$$

Note that as $N \to \infty$, we expect the posterior prob. to have vanishing variance: $S_N \to 0$ as $N \to \infty$.

Indeed,

$$(S_N^{-1})_{ij} = \alpha I + \beta (\Phi^T \Phi)_{ij} =$$

$$= \alpha I + \beta \sum_{n=1}^{N} \varphi_i(\vec{x}_n) \varphi_j(\vec{x}_n) .$$

This $\nearrow$ as $N \uparrow \Rightarrow S_N \searrow$.

Moreover, if $\vec{\varphi}(\vec{x})$ is localized,

$\sigma_N^2(\vec{x}) \to \beta^{-1}$ even for small N, if $\vec{x}$ is far away from the basis function centers. Thus, the model may become overconfident when extrapolating.