Machine learning algorithms

Schedule
next week

Tomorrow:
async.
- posted
short video
Lasso
- jupyter

Probability & priors

2020–10-12

Late HW:
2 days

CSCI 471 / 571, Fall 2020

Kameron Decker Harris

Ridge regression 3

- Way to control bias-variance tradeoff
- Regularization $\text{loss} + \lambda \|\vec{\beta}\|^2$
  - Hyperparameter λ ← controls strength
  - Shrinks coefficients

Practical considerations: Ridge

- Best if features X are standardized

$$\begin{bmatrix} 10^6 & 1 \\ 10^5 & 3 \\ \vdots & \end{bmatrix} \rightarrow \begin{bmatrix} 3 & -1 \\ .3 & 0.3 \end{bmatrix}$$

subtract off mean, divide by standard deviation

j^{th} coordinate

i^{th} example

$$X_{ij} \longrightarrow \frac{X_{ij} - \mu_j}{\sigma_j}$$

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

$$\sigma_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \mu_j)^2$$

- Don't penalize intercept

$$f(\vec{x}) = \underline{\underline{\beta_0}} + \sum_{i=1}^{d} \beta_i x_i$$

helps fit mean of $\vec{y}$

$$\lambda \sum_{i=1}^{d} \beta_i^2$$

$$\cancel{\lambda \sum_{i=0}^{d} \beta_i^2}$$

Probability and priors

pixabay

in Oct
likely, impossible

in August
less likely

know from prior experience

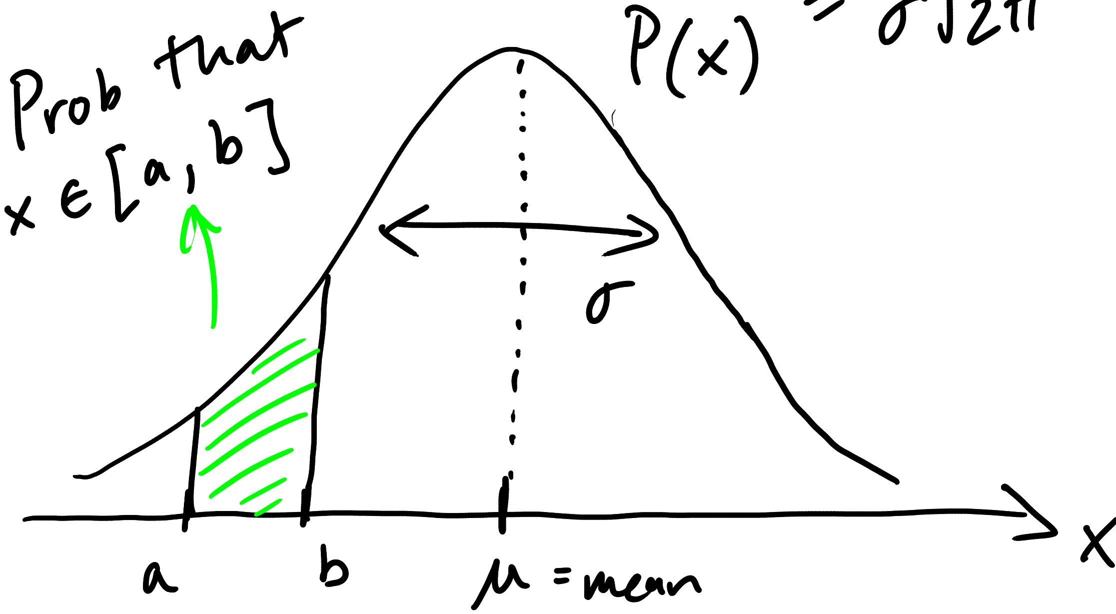
rainy days

Basic probability: normal distribution

Gaussian, bell-shaped

$X \sim \mathcal{N}(\mu, \sigma^2)$

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$



mean or <u>expectation</u>

$$\begin{cases} \mathbb{E}[X] = \mu \\ \mathrm{Var}[X] \end{cases}$$

how much it <u>varies</u>

$$= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

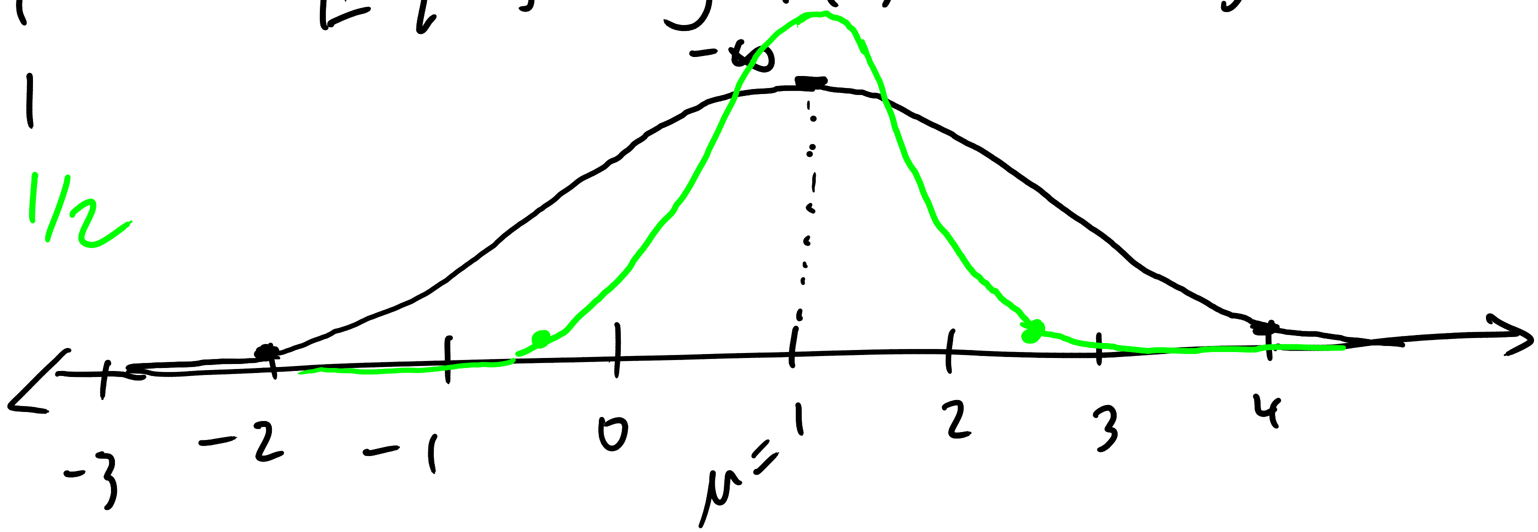
$$= \sigma^2$$

ex/ $\mu = 1$

$\sigma = 1$

$\sigma = 1/2$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} P(x)\, x \, dx = \mu$$



$3\sigma = 3$

estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \, , \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Bayes' rule

Allows you to incorporate prior knowledge.
Update prior w/ new information

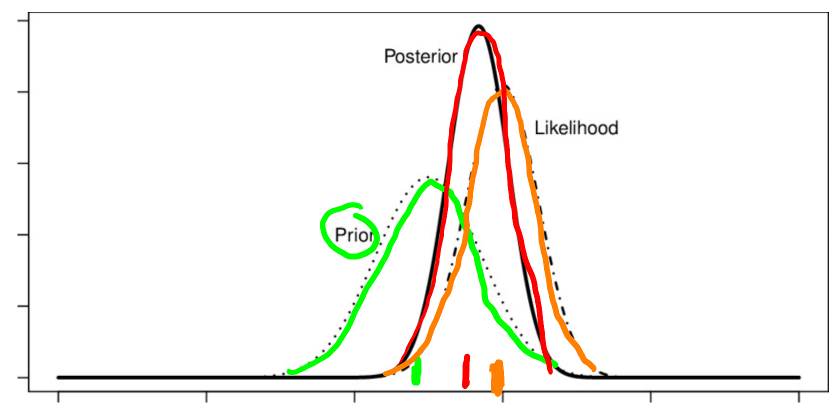
A, B $\quad A =$ will it rain? $\quad B =$ whether today it rained

Bayes' rule

$$\underbrace{\boxed{P(A|B)}}_{\text{posterior}} = \frac{\overbrace{\boxed{P(B|A)}}^{\text{likelihood}}\overbrace{\boxed{P(A)}}^{\text{prior}}}{P(B)}$$

$P(B) \leftarrow$ marginal distribution

$A =$ parameters, e.g. $\vec{\beta}$
$B =$ data $\vec{y}$



https://www.researchgate.net/figure/Bayesian-updating-of-the-prior-distribution-to-posterior-distribution-The-Posterior_fig1_320507985

MAP estimator

$B =$ data $\vec{y}$, X

$A =$ parameters $\vec{\beta}$

Assume: $y_i = \vec{x}_i^T \vec{\beta} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ independent identically distributed

Ridge is equivalent to "max a posteriori"

$$P(\vec{y} \mid \vec{\beta}) \overset{\text{(indep)}}{=} P(y_1 \mid \vec{\beta})\, P(y_2 \mid \vec{\beta}) \cdots P(y_n \mid \vec{\beta})$$

likelihood

$$= \prod_{i=1}^{n} P(y_i \mid \vec{\beta})$$

maximum likelihood = linear regression OLS

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\vec{x}_i^T\vec{\beta} - y_i}{\sigma}\right)^2\right)$$

$$= C \cdot \exp\left(-\frac{1}{2\sigma^2} \| X\vec{\beta} - \vec{y} \|^2\right)$$

likelihood ✓

prior : $\underline{\vec{\beta} \sim N\left(0, \frac{1}{\lambda}\right)}$,

$$P(\vec{\beta}) = \prod_{i=1}^{d} \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda}{2} \beta_i^2\right)$$

$$= \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-\frac{\lambda}{2} \underbrace{\|\vec{\beta}\|^2}_{\text{Ridge}}\right)$$

Use Bayes' rule

$$P(\vec{y} \mid \vec{\beta})\, P(\vec{\beta}) = C \cdot \exp\left(-\frac{1}{2\sigma^2} \|X\vec{\beta} - \vec{y}\|^2\right) \exp\left(-\frac{\lambda}{2} \|\vec{\beta}\|^2\right)$$

$$\max_{\vec{\beta}} P(\vec{\beta} \mid \vec{y})$$

$$= \max_{\vec{\beta}} \underset{\text{monotonic}}{\log} P(\vec{\beta} \mid \vec{y}) = \max_{\vec{\beta}} \log C - \frac{1}{2\sigma^2} \|X\vec{\beta} - \vec{y}\|^2 - \frac{\lambda}{2} \|\vec{\beta}\|^2$$

(− cost of Ridge)