Maximum Likelihood, MMSE

- maximum likelihood estimation
- regression
- MMSE

Big Picture

- machine learning is about **learning functions from data**

$$\widehat{y} = f(\boldsymbol{x})$$

- the inputs to the functions are called **features**
- the true output is called a label

$\boldsymbol{x} =$ 

classification: $y \in \{0, 1, \ldots, 9\}$ regression: $y \in \mathbb{R}$

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \qquad (\boldsymbol{x}_i, y_i) \overset{i.i.d.}{\sim} p(\boldsymbol{x}, y)$$

- if we have a good approximation of $p(\boldsymbol{x}, y)$, we know what to do:

$$\widehat{y} = \arg\max_y p(y|\boldsymbol{x})$$ MAP estimate

$$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)p(y)$$ posterior $\propto$ likelihood $\times$ prior

$$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)$$ ML (maximum likelihood) estimate

$$\widehat{y} = \arg\max_y p_y(\boldsymbol{x})$$

Maximum Likelihood Example

$$X_i \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$$

$$p(x) = \begin{cases} \theta & x = 1 \\ (1-\theta) & x = 0 \end{cases}$$

- flip a coin with unknown bias n times $\quad H, H, H, T$
- how can we estimate bias of the coin θ?

$$p(\boldsymbol{x}) = p(x_1 = 1)p(x_2 = 1)p(x_3 = 1)p(x_4 = 0)$$

$$p_\theta(\boldsymbol{x}) = \theta^3(1-\theta)$$

- think of $p_\theta(\boldsymbol{x})$ as a family of distributions parameterized by θ

$$p_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} p_\theta(x_i)$$

$$p_\theta(\boldsymbol{x}) = \theta^k(1-\theta)^{n-k}$$

where k is number of heads

- answer should be $\widehat{\theta} = \frac{k}{n}$

Maximum Likelihood

$$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)p(y)$$

$$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)$$

$$\widehat{y} = \arg\max_y p_y(\boldsymbol{x})$$

$$\widehat{\theta} = \arg\max_\theta p_\theta(\boldsymbol{x})$$

- observe some data $\boldsymbol{x}$
- family of distributions parametrized by θ

$$p_\theta(\boldsymbol{x})$$

$$\widehat{\theta} = \arg\max_\theta p_\theta(\boldsymbol{x})$$

- $L(\theta) = p_\theta(\boldsymbol{x})$ is the likelihood function (fixed $\boldsymbol{x}$, function of θ)
- $\ell(\theta) = \log(p_\theta(\boldsymbol{x}))$ is the log-likelihood function (fixed $\boldsymbol{x}$, function of θ)

$$\widehat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$$

$\widehat{\theta}$ is the maximum likelihood estimate of θ

Maximum Likelihood for Mean of Gaussian

- model: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$ for unknown θ

$$L(\theta) = p_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} p_\theta(x_i) = \prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}$$

'likelihood of θ given we saw the data $\boldsymbol{x}$'

'what's the most likely θ given we saw the data $\boldsymbol{x}$?'

$$\arg\max_\theta L(\theta) = \arg\max_\theta \prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}$$

```
x = np.random.randn(100,1) + mu
print(x)
```

$$\boldsymbol{x} = \begin{matrix}[[-0.10223044] \\ [\ 0.74557181] \\ [\ 1.61851936] \\ [-1.05609849] \\ [-0.50006279] \\ [-1.24555422] \\ \ldots\end{matrix}$$

$$= \arg\min_\theta \sum_i (x_i - \theta)^2$$

$$\widehat{\theta}_{\mathrm{ML}} = \frac{1}{n} \sum_i x_i$$

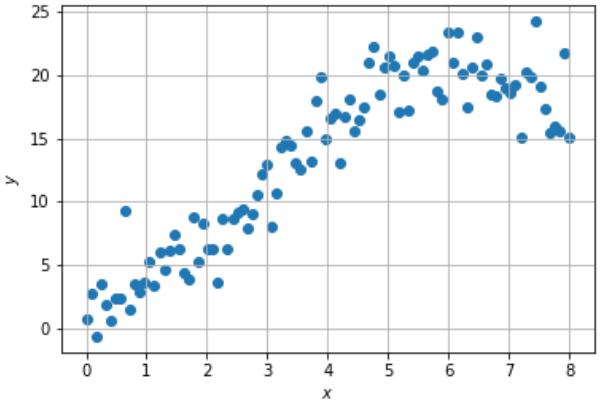
Regression and Maximum Likelihood

- model y as a weighted sum of features $\boldsymbol{x}$, plus noise Z

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + Z_i \qquad Z_i \sim \mathcal{N}(0, \sigma^2)$$

$$y_i \sim \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}_i, \sigma^2)$$

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$$



- what's the likelihood of $\boldsymbol{w}$ given we saw $\mathcal{D}$?

$$L(\boldsymbol{w}) = p_{\boldsymbol{w}}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}}$$

- goal: find coefficients $\boldsymbol{w}$ so that

$$y = w_1 x^3 + w_2 x^2 + w_3 x^1 + w_4$$

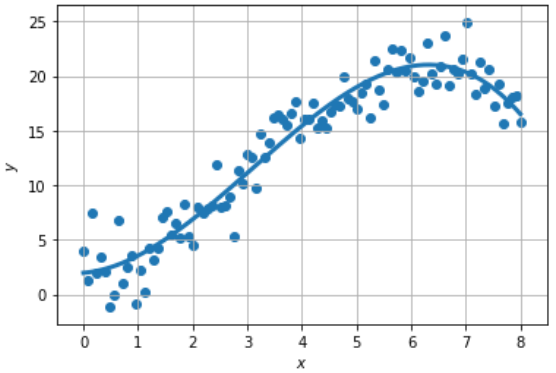
$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

$$\boldsymbol{x}_i^T = [x_i^3 \quad x_i^2 \quad x_i^1 \quad 1]$$

Regression and Maximum Likelihood

$$L(\boldsymbol{w}) = p_{\boldsymbol{w}}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{w})^2}{2\sigma^2}}$$

$$\boldsymbol{w}_{\mathrm{ML}} = \arg\max_{\boldsymbol{w}} L(\boldsymbol{w})$$



$$= \arg\min_{\boldsymbol{w}} \sum_i (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

$$= \arg\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2$$

$$\boldsymbol{x}^T = [x^3 \quad x^2 \quad x^1 \quad 1]$$

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

- maximum likelihood with gaussian noise gives rise to the squared error loss

Minimum Mean Square Error Estimation

Minimum Mean Squared Error

recap . . .

- MAP estimate: $\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)p(y)$
- ML estimate: $\widehat{y} = \arg\max_y p_y(\boldsymbol{x})$
    - **claim**: MAP minimizes risk under 0/1 loss.
    - **claim**: ML minimizes risk under 0/1 loss and uniform prior.

$$E[\ell(\widehat{y}, y)] = E[\mathbb{1}_{\{\widehat{y} \neq y\}}] = \mathbb{P}(\widehat{y} \neq y)$$

MAP by definition picks the most probable y.

- MMSE (Minimum mean squared error) estimate:

$$E[\ell(\widehat{y}, y)] = E[(\widehat{y} - y)^2]$$

- **claim (obvious)**: MMSE minimizes risk under squared error loss.

Minimum Mean Squared Error Estimation

- we observe features $\boldsymbol{x}$ and want to estimate y, and we know $p(\boldsymbol{x}, y)$
- find an estimate $\widehat{y}$ that minimizes risk with squared error loss:

$$E[\ell(y, \widehat{y})] \ = E[(y - \widehat{y})^2]$$

- claim: the MMSE estimate of y given the features $\boldsymbol{x}$ and complete knowledge of $p(\boldsymbol{x}, y)$ is:

$$\widehat{y} = E[y|\boldsymbol{x}]$$

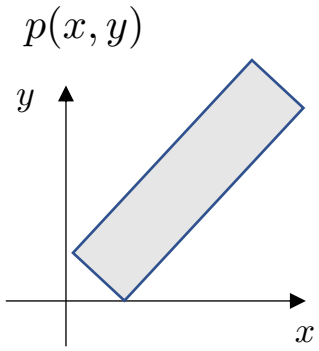
- note: MMSE is basis for denoising techniques such as Wiener filter, Kalman Filter.

MMSE Example

- claim: the MMSE estimate of y given the features $\boldsymbol{x}$ and complete knowledge of $p(\boldsymbol{x}, y)$ is:

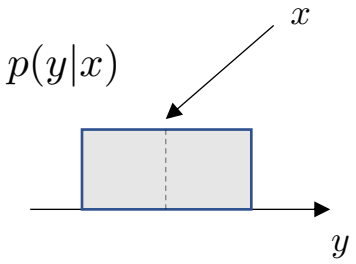
$$\widehat{y} = f(\boldsymbol{x}) = E[y|\boldsymbol{x}]$$

- example:



- given x, what is your best guess for y?

$$y|x \sim U[x - \Delta, x + \Delta]$$



- MMSE: $\widehat{y} = E[y|x] = x$
- MAP: $\widehat{y}$ can be anything between $x - \Delta$ and $x + \Delta$

MMSE Proof

- claim: the MMSE estimate of y given the features $\boldsymbol{x}$ and complete knowledge of $p(\boldsymbol{x}, y)$ is:

$$\widehat{y} = f(\boldsymbol{x}) = E[y|\boldsymbol{x}]$$

- proof:

1. note that $\min_a E[(y-a)^2]$ is given by $a = E[y]$ since

$$\begin{aligned} E[(y-a)^2] &= E[((y-E[y]) + (E[y]-a))^2] \\ &= E[((y-E[y])^2 + (E[y]-a)^2 + 2(y-E[y])(E[y]-a)] \\ &= E[(y-E[y])^2] + (E[y]-a)^2 \\ &\geq E[(y-E[y])^2] \quad \text{with equality if and only if } E[y] = a \end{aligned}$$

2. for fixed $\boldsymbol{x}$, we have that $E_y[(y-f(\boldsymbol{x}))^2|\boldsymbol{x}]$ is minimized by $f(\boldsymbol{x}) = E[y|\boldsymbol{x}]$

$$E[(y-f(\boldsymbol{x}))^2] = E_{\boldsymbol{x}}[E_y[(y-f(\boldsymbol{x}))^2|\boldsymbol{x}]]$$

$f(\boldsymbol{x}) = E[y|\boldsymbol{x}]$ minimizes mean squared error.

MMSE

- claim: the minimum risk under the squared error loss given the features $\boldsymbol{x}$

  and complete knowledge of $p(\boldsymbol{x}, y)$ is:

  $E_{\boldsymbol{x}}[\mathrm{var}(y|\boldsymbol{x})]$

- the MMSE (the error, not the estimate) of y given the features $\boldsymbol{x}$ and complete knowledge of $p(\boldsymbol{x}, y)$ is:

  $E_{\boldsymbol{x}}[\mathrm{var}(y|\boldsymbol{x})]$

- proof: