Lecture 10/3: Regularized Linear Models

Thursday, October 3, 2019 11:03 AM

Bayes Decision Rule

Example: x describes patient, y whether drug will cure patient

Know $P(y=1 \mid x=x) = 0.7$

$P(y=-1 \mid x=x) = 0.3$

predicting $y=1$: make error with probability 0.3

$y=-1$: make error with probability 0.7

→ predict label y that has the highest $P(y \mid x=x)$

Bayes Error Rate

What is $err_p(h_{bayes})$ for 0/1 loss?

Given an instance x, h_{bayes} has probability of error of $\min_{y \in Y} (1 - P(y=y \mid x=x))$

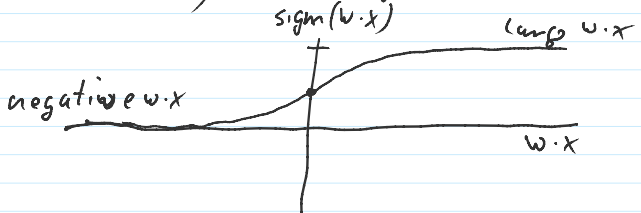
→ Expectation over all $x \sim P(x)$:

$$err_p(h_{bayes}) = \sum_{x \in X} P(x=x) \min_{y \in Y} (1 - P(y=y \mid x=x))$$

Logistic Regression

Idea: Model $P(y \mid x) = Ber(y \mid \mu = sigm(w \cdot x))$ for binary classification

sigmoid: $sigm(w \cdot x) = \frac{1}{1+e^{-w \cdot x}}$



$P(y=1 \mid x, w) = sigm(w \cdot x) = \frac{1}{1+e^{-w \cdot x}}$

$P(y=-1 \mid x, w) = 1 - sigm(w \cdot x) = 1 - \frac{1}{1+e^{-w \cdot x}} = \frac{1+e^{-w \cdot x} - 1}{1+e^{-w \cdot x}} = \frac{1}{e^{w \cdot x}(1+e^{-w \cdot x})}$

$= \frac{1}{1+e^{w \cdot x}}$

→ $P(y \mid x, w) = \frac{1}{1+e^{-y\, w \cdot x}}$

Training: Conditional Maximum Likelihood

$$w = \arg\max_w P(y_1 \dots y_m \mid x_1 \dots x_m, w)$$

$$= \arg\max_w \prod_{i=1}^{m} P(y_i \mid x_i, w)$$

$$= \operatorname{argmax}_w \prod_{i=1}^{m} P(y_i \mid x_i, w)$$

$$= \operatorname{argmax}_w \prod_{i=1}^{m} \frac{1}{1+e^{-y_i w \cdot x_i}}$$

$$= \operatorname{argmin} -\log \left[\prod_{i=1}^{m} \frac{1}{1+e^{-y_i w \cdot x_i}} \right]$$

$$= \operatorname{argmin} \sum_{i=1}^{m} -\log \left(\frac{1}{1+e^{-y_i w \cdot x_i}} \right)$$

$$= \operatorname{argmin}_w \sum_{i=1}^{m} \log \left(1+e^{-y_i w \cdot x_i} \right)$$

Regularized Logistic Regression

Idea: Like Logistic Regression, but put prior $P(w)$ on the parameters

$\rightarrow P(w) = N(w \mid 0, \operatorname{diag}(\sigma)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{w \cdot w}{2\sigma^2}}$

Training: Maximum a posteriori estimate

$$w = \operatorname{argmax}_w P(y_1 \dots y_m \mid x_1 \dots x_m, w) \cdot P(w)$$

$$= \operatorname{argmin}_w \left[\sum_{i=1}^{m} \log \left(1+e^{-y_i w \cdot x_i} \right) \right] - \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \log \left(e^{-\frac{w \cdot w}{2\sigma^2}} \right)$$

$$= \operatorname{argmin}_w \frac{1}{2} w \cdot w + \sigma^2 \sum_{i=1}^{m} \log \left(1+e^{-y_i w \cdot x_i} \right)$$

$\rightarrow$ SVM Primal: $w = \operatorname{argmin} \frac{1}{2} w \cdot w + C \sum_{i=1}^{m} \max(0, 1 - y_i w \cdot x_i)$

Ridge Regression

Regression: Predict real-valued $y \in \mathbb{R}$ from features $x \in \mathbb{R}^N$

Idea: Model likelihood $P(y \mid x, w) = N(y \mid w \cdot x, \varepsilon) = \frac{1}{Z} e^{\frac{-(w \cdot x - y)^2}{2\varepsilon^2}}$

and prior $P(w) = N(w \mid 0, \operatorname{diag}(\sigma))$

Training: Max a posteriori

$$w = \operatorname{argmax}_w \left[\prod_{i=1}^{m} P(y_i \mid x_i, w) \right] \cdot P(w)$$

$$= \operatorname{argmin}_w -\left[\sum_{i=1}^{m} \log P(y_i \mid x_i, w) \right] - \log P(w)$$

$$= \operatorname{argmin} \left[\frac{1}{} \sum_{} (\dots \cdot \dots)^2 \right] - 1 \dots$$

$$= \arg\min_w \left[\frac{1}{2\varepsilon^2} \sum_{i=1}^{m} (w \cdot x_i - y_i)^2 \right] - \frac{1}{2\sigma^2} w \cdot w$$

$$= \arg\min_w \frac{1}{2} w \cdot w + \frac{\sigma^2}{2\varepsilon^2} \sum_{i=1}^{m} (w \cdot x_i - y_i)^2$$

Regularized linear Models

L:

- Hinge loss: $\max(0, 1 - y\bar{y})$
  Exp loss: $\exp(-\beta y \cdot \bar{y})$
  Squared error: $(y - \bar{y})^2$
  Absolute error: $|y - \bar{y}|$

R:

L_2 norm: $\|w\|_2 = \sqrt{w \cdot w}$

L_1 norm: $\|w\|_1 = \sum |w_i|$

L_0 norm: $\|w\|_0$

L_p norm

Beyond linear models:

linear: $w \cdot x + b$

Deep network: $f_w(x)$

Kernel: $w \cdot \phi(x) + b$