Machine learning: lecture 6

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Topics

- Regularization
  - prior, penalties, MAP estimation
  - the effect of regularization, generalization
  - regularization and discrimination
- Discriminative classification
  - criterion, margin
  - support vector machine

MAP estimation, regularization

- Consider again a simple 2-d logistic regression model

$$P(y=1|\mathbf{x},\mathbf{w}) = g\left(w_0 + w_1 x_1 + w_2 x_2 \right)$$

- Before seeing any data we may prefer some values of the parameters over others (e.g., small over large values).

MAP estimation, regularization

- Consider again a simple 2-d logistic regression model

$$P(y=1|\mathbf{x},\mathbf{w}) = g\left(w_0 + w_1 x_1 + w_2 x_2 \right)$$

- Before seeing any data we may prefer some values of the parameters over others (e.g., small over large values).
- We can express this preference through a prior distribution over the parameters (here omitting w_0)

$$p(w_1, w_2; \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}(w_1^2 + w_2^2) \right\}$$

where σ^2 determines how tightly around zero we want to constrain the values of w_1 and w_2.

MAP estimation, regularization

- Consider again a simple 2-d logistic regression model

$$P(y=1|\mathbf{x},\mathbf{w}) = g\left(w_0 + w_1 x_1 + w_2 x_2 \right)$$

- Before seeing any data we may prefer some values of the parameters over others (e.g., small over large values).
- We can express this preference through a prior distribution over the parameters (here omitting w_0)

$$p(w_1, w_2; \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}(w_1^2 + w_2^2) \right\}$$

- To combine the prior with the availabale data we find the MAP (maximum a posteriori) parameter estimates:

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i,\mathbf{w}) \right] p(w_1, w_2; \sigma^2)$$

MAP estimation, regularization

- The estimation criterion is now given by a *penalized* log-likelihood (cf. log-posterior):

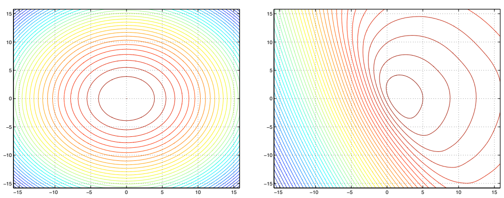
$$\begin{aligned} \tilde{l}(D;\mathbf{w}) &= \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) + \log p(w_1, w_2; \sigma^2) \\ &= \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}(w_1^2 + w_2^2) + \text{const.} \end{aligned}$$

- We'd like to understand how the solution changes as a function of the prior variance σ^2 (or more generally with different priors)

The effect of regularization

- Let's first understand graphically how the addition of the prior changes the solution

$$\tilde{l}(D;\mathbf{w}) \;=\; \overbrace{\sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w})}^{\text{log-likelihood}} \overbrace{-\frac{1}{2\sigma^2}(w_1^2+w_2^2)}^{\text{log-prior}} + \text{const.}$$

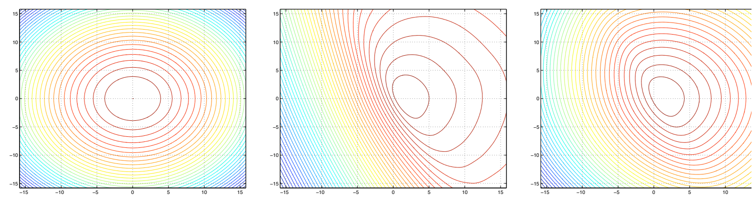


log-prior log-likelihood

The effect of regularization

- Let's first understand graphically how the addition of the prior changes the solution

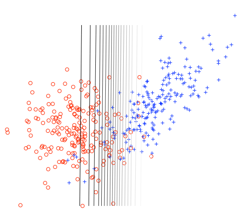
$$\tilde{l}(D;\mathbf{w}) \;=\; \overbrace{\sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w})}^{\text{log-likelihood}} \overbrace{-\frac{1}{2\sigma^2}(w_1^2+w_2^2)}^{\text{log-prior}} + \text{const.}$$



log-prior log-likelihood log-posterior

The effect of regularization cont'd

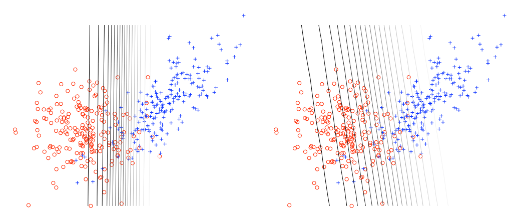
$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}(w_1^2+w_2^2) + \text{const.}$$



$\sigma^2 = \infty$

The effect of regularization cont'd

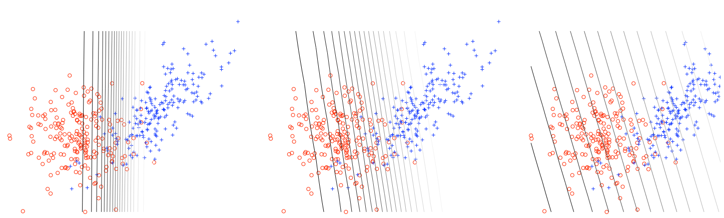
$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}(w_1^2+w_2^2) + \text{const.}$$



$\sigma^2 = \infty$ $\sigma^2 = 10/n$

The effect of regularization cont'd

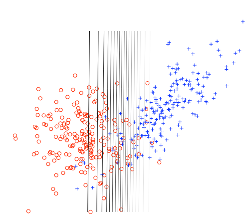
$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}(w_1^2+w_2^2) + \text{const.}$$



$\sigma^2 = \infty$ $\sigma^2 = 10/n$ $\sigma^2 = 1/n$

The effect of regularization cont'd

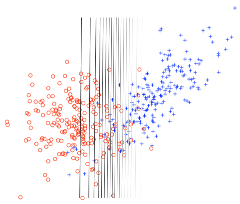
$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}w_1^2 + \text{const.}$$



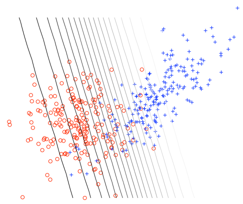
$\sigma^2 = \infty$

The effect of regularization cont'd

$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}w_1^2 + \text{const.}$$



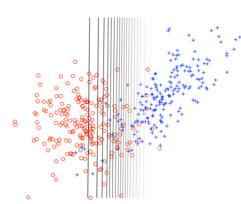
$\sigma^2 = \infty$



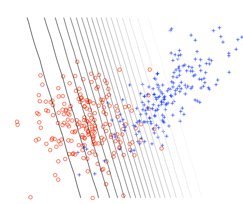
$\sigma^2 = 10/n$

The effect of regularization cont'd

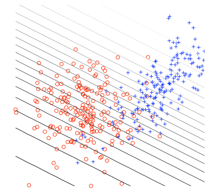
$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}w_1^2 + \text{const.}$$



$\sigma^2 = \infty$



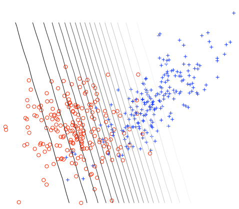
$\sigma^2 = 10/n$



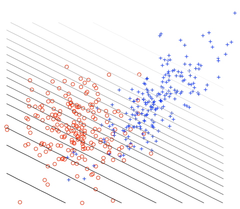
$\sigma^2 = 1/n$

The effect of regularization cont'd

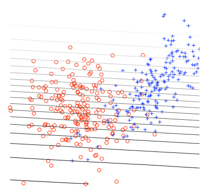
$$\tilde{l}(D;\mathbf{w}) \;=\; \sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2}w_1^2 + \text{const.}$$



$\sigma^2 = 10/n$



$\sigma^2 = 1/n$



$\sigma^2 = 0.1/n$

The effect of regularization: train/test

- (Scaled) penalized log-likelihood criterion

$$\tilde{l}(D;\mathbf{w})/n \;=\; \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{n2\sigma^2}(w_1^2+w_2^2) + \text{const.}$$

The effect of regularization: train/test

- (Scaled) penalized log-likelihood criterion

$$\tilde{l}(D;\mathbf{w})/n \;=\; \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{c}{2}(w_1^2+w_2^2) + \text{const.}$$

where $c = 1/n\sigma^2$; increasing c results in stronger regularization.

The effect of regularization: train/test

- (Scaled) penalized log-likelihood criterion

$$\tilde{l}(D;\mathbf{w})/n \;=\; \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{c}{2}(w_1^2+w_2^2) + \text{const.}$$

where $c = 1/n\sigma^2$; increasing c results in stronger regularization.

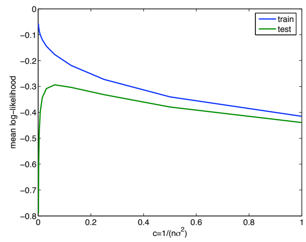
- Resulting average log-likelihoods

$$\text{training log-lik.} \;=\; \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\hat{\mathbf{w}}_{MAP})$$

$$\text{test log-lik.} \;=\; E_{(\mathbf{x},y)\sim P}\left\{\log P(y|\mathbf{x},\hat{\mathbf{w}}_{MAP})\right\}$$

The effect of regularization: train/test

$$\begin{aligned} \tilde{l}(D;\mathbf{w})/n &= \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{c}{2}(w_1^2+w_2^2) + \text{const.} \\ \text{training log-lik.} &= \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\hat{\mathbf{w}}_{MAP}) \\ \text{test log-lik.} &= E_{(\mathbf{x},y)\sim P}\{\log P(y|\mathbf{x},\hat{\mathbf{w}}_{MAP})\} \end{aligned}$$

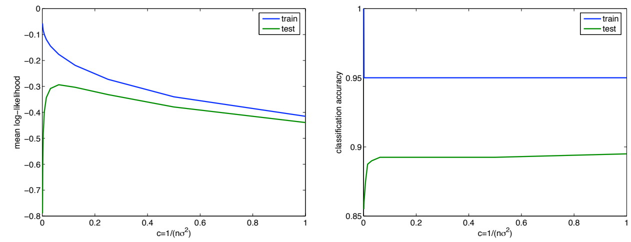


Likelihood, regularization, and discrimination

- Regularization by penalizing $\|\mathbf{w}_1\|^2 = w_1^2 + w_2^2$ in

$$\tilde{l}(D;\mathbf{w})/n = \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{c}{2}(w_1^2+w_2^2) + \text{const.}$$

does not *directly* limit the logistic regression model as a classifier. For example:



Likelihood, regularization, and discrimination

- Regularization by penalizing $\|\mathbf{w}_1\|^2 = w_1^2 + w_2^2$ in

$$\tilde{l}(D;\mathbf{w})/n = \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{c}{2}(w_1^2+w_2^2) + \text{const.}$$

does not *directly* limit the logistic regression model as a classifier.

- Classification decisions only depend on the sign of the *discriminant function*

$$f(\mathbf{x};\mathbf{w}) = w_0 + \mathbf{x}^T\mathbf{w}_1 = (\mathbf{x}-\mathbf{x}_0)^T\mathbf{w}_1$$

where $\mathbf{w}_1 = [w_1, w_2]^T$ and $\mathbf{x}_0$ is chosen such that $w_0 = \mathbf{x}_0^T\mathbf{w}_1$. Limiting $\|\mathbf{w}_1\|^2 = w_1^2 + w_2^2$ does not reduce the possible signs.

Topics

- Regularization
  - prior, penalties, MAP estimation
  - the effect of regularization, generalization
  - regularization and discrimination
- Discriminative classification
  - criterion, margin
  - support vector machine

Discriminative classification

- Consider again a binary classification task with $y = \pm 1$ labels (not 0/1 as before) and linear discriminant functions

$$f(\mathbf{x};\mathbf{w}) = w_0 + \mathbf{x}^T\mathbf{w}_1$$

parameterized by w_0 and $\mathbf{w}_1 = [w_1, \ldots, w_d]^T$.

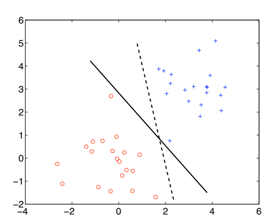
- The predicted label is simply given by the sign of the discriminant function $\hat{y} = \text{sign}(f(\mathbf{x};\mathbf{w}))$
- We are only interested in getting the labels correct; no probabilities are associated with the predictions

Discriminative classification

- When the training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is *linearly separable* we can find parameters $\mathbf{w}$ such that

$$y_i[w_0 + \mathbf{x}_i^T\mathbf{w}_1] > 0,\ i = 1, \ldots, n$$

i.e., the sign of the discriminant function agrees with the label

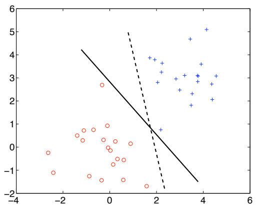


(there are many possible solutions)

Discriminative classification

- Perhaps we can find a better discriminant boundary by requiring that the training examples are separated with a fixed "margin":

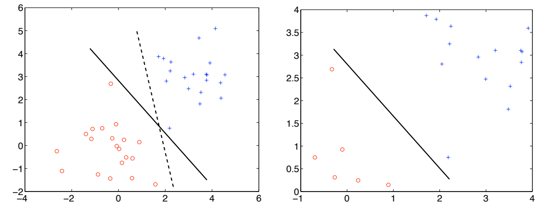
$$y_i[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \; i = 1, \ldots, n$$



Discriminative classification

- Perhaps we can find a better discriminant boundary by requiring that the training examples are separated with a fixed "margin":

$$y_i[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \; i = 1, \ldots, n$$



The problem is the same as before. The notion of "margin" used here depends on the scale of $\|\mathbf{w}_1\|$

Margin and regularization

- We get a more meaningful (geometric) notion of margin by regularizing the problem:

$$\text{minimize} \;\; \frac{1}{2}\|\mathbf{w}_1\|^2 = \frac{1}{2}\sum_{i=1}^{d} w_i^2$$

subject to

$$y_i[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \;\; i = 1, \ldots, n$$

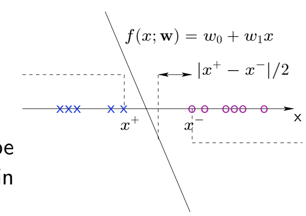
- What can we say about the solution?

Margin and regularization

- One dimensional example: $f(x; \mathbf{w}) = w_0 + w_1 x$

Relevant constraints:

$$\begin{aligned} 1[w_0 + w_1 x^+] - 1 &\geq 0 \\ -1[w_0 + w_1 x^-] - 1 &\geq 0 \end{aligned}$$



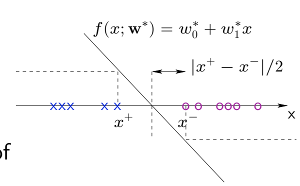
Maximum separation would be at the mid point with a margin $|x^+ - x^-|/2$.

Margin and regularization

- One dimensional example: $f(x; \mathbf{w}) = w_0 + w_1 x$

Relevant constraints:

$$\begin{aligned} 1[w_0 + w_1 x^+] - 1 &\geq 0 \\ -1[w_0 + w_1 x^-] - 1 &\geq 0 \end{aligned}$$



At the mid point the value of the margin is $|x^+ - x^-|/2$.

- We can find the maximum margin solution by minimizing the slope $|w_1|$ while satisfying the classification constraints
- The resulting margin is directly tied to the minimizing slope (slope = 1/margin): $|w_1^*| = 2/|x^+ - x^-|$

Support vector machine

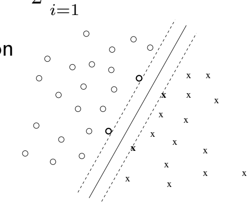
- We minimize the regularization penalty

$$\frac{1}{2}\|\mathbf{w}_1\|^2 = \frac{1}{2}\sum_{i=1}^{d} w_i^2$$

subject to the classification constraints

$$y_i[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0$$

for $i = 1, \ldots, n$.



- Analogously to the one dimensional case, the "slope" is related to the geometric margin: $\|\mathbf{w}_1^*\| = 1/\text{margin}$.
- The solution is again defined only on the basis of a subset of examples or "support vectors"