Classification

Probabilistic Discriminative Models II

T.L. Grobler
(with some edits by S. Kroon)

Classification

- Given $\mathbf{x}$ assign to one of k classes:
  - C_j, $j = 1,\ldots,k$
  - Assign prob $P(C_j|\mathbf{x})$
  - $C^* = \mathrm{argmax}_{C_j} P(C_j|\mathbf{x})$
- Class prob, more useful than knowing max class prob.

Discriminative Approach

- Dispenses with: $p(\mathbf{x}|C_j)$
- Directly compute posterior $P(C_j|\mathbf{x})$
- Probabilistic Discriminative Models (PDM)

Problems

- Weights become too large
  - $P(C_1|\mathbf{x}_n,\mathbf{w}) \rightarrow 1$
  - $\mathbf{w}^T\mathbf{x}_n \rightarrow \infty$
- Overfitting
  - Samples at boundary have large influence
  - Boundary from training data
    - Fails to generalize
    - Too specific

Solutions

- Constrained optimization
  - $\mathbf{w}^T\mathbf{w} = 1$
- Add penalty term
  - regularization

Bayesian Approach

- **w** a parameter:
  - Prev approach
  - MLE: $\mathbf{w}^* \rightarrow P(C_1|\mathbf{x},\mathbf{w}^*)$
- **w** a random variable:
  - Posterior Class Probability $P(C_1|\mathbf{x},D)$?
  - D training data
  - **x** observation

Posterior Class Probability

Marginalization

$$P(C_1|\mathbf{x}, D) = \int P(C_1, \mathbf{w}|\mathbf{x}, D)d\mathbf{w}$$

$$= \int P(C_1|\mathbf{w}, \mathbf{x}, D)p(\mathbf{w}|\mathbf{x}, D)d\mathbf{w}$$

$$= \int P(C_1|\mathbf{w}, \mathbf{x})p(\mathbf{w}|D)d\mathbf{w}$$

Independence x 2

Prob Chain Rule

Integral Evaluation

- Markov Chain Monte Carlo (MCMC)
  - Averaging
  - Marginalization
- MAP (Maximum A Posteriori)
  - Approximate integral
  - $p(\mathbf{w}|D)$ sharply peaked at mode $\mathbf{w}^*$
  - $\int P(C_1|\mathbf{w},\mathbf{x})p(\mathbf{w}|D)d\mathbf{w} \approx P(C_1|\mathbf{w}^*,\mathbf{x})$

~ delta function - $\delta(\mathbf{w}^*)$

MAP Estimate

$$
\begin{aligned}
E(\mathbf{w}) &= -\ln p(\mathbf{w}|D) \\
\mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}}\ E(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}}\ -\ln p(\mathbf{w}|D) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}}\ -\ln p(D|\mathbf{w})p(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}}\ -\ln p(D|\mathbf{w}) - \ln p(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}}\ l(\mathbf{w})
\end{aligned} \tag{1}
$$

Bayes Rule

Prior?

$$\underset{\mathbf{w}}{\operatorname{argmin}} \; -\ln p(D|\mathbf{w}) - \ln p(\mathbf{w})$$

$$p(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \lambda\mathbf{I})$$

$$\ln p(\mathbf{w}) = -\frac{1}{2\lambda}\mathbf{w}^T\mathbf{w}$$

Isotropic Normal

$$\sum_{n=1}^{N}\{y_n \ln\sigma(\mathbf{w}^T\mathbf{x}_n) + (1-y_n)\ln(1-\sigma(\mathbf{w}^T\mathbf{x}_n))\} - \ln P(X)$$

MAP v MLE

- Difference
  - Regularization term
  - $(2\lambda)^{-1}\mathbf{w}^T\mathbf{w}$
- Frequentist
  - Realizes necessity of regularization term
  - Prevents overfitting of MLE
- Bayesian
  - Penalty appears naturally in MAP as log-prior

Hyperparameter

- Regularization term
  - Depends on hyperparameter λ
  - Determined by validation set

Newton-Raphson

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \; -\ln p(D|\mathbf{w}) - \ln p(\mathbf{w})$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} \; l(\mathbf{w})$$

$$\boldsymbol{\Delta} l = \mathbf{0}$$

$$\boxed{\mathbf{w}_{k+1} = \mathbf{w}_k - \mathbf{H}^{-1}\boldsymbol{\Delta} l}$$

Newton-Raphson

$$\boldsymbol{\Delta} l = -\sum_{n=1}^{N}(y_n - \sigma(\mathbf{w}^T\mathbf{x}_n))\mathbf{x}_n + \frac{1}{\lambda}\mathbf{w}$$

$$\mathbf{H} = \sum_{n=1}^{N}\sigma(\mathbf{w}^T\mathbf{x}_n)(1 - \sigma(\mathbf{w}^T\mathbf{x}_n))\mathbf{x}_n\mathbf{x}_n^T + \frac{1}{\lambda}\mathbf{I}$$

Hessian Matrix

Hessian Positive Definite

$$
\begin{aligned}
\mathbf{z}^T\mathbf{H}\mathbf{z} &= \sum_{n=1}^{N} \sigma_n(1-\sigma_n)\mathbf{z}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{z} + \frac{1}{\lambda}\|\mathbf{z}\|^2 \\
&= \sum_{n=1}^{N} \sigma_n(1-\sigma_n)\|\mathbf{x}_n^T\mathbf{z}\|^2 + \frac{1}{\lambda}\|\mathbf{z}\|^2 \\
&> 0
\end{aligned}
$$

w* global minimum

Multi-class Logistic Regression

$$P(C_i|\mathbf{x}, \mathbf{w}_i) = \frac{\exp\left(\mathbf{w}_i^T \mathbf{x}\right)}{\sum_{j=1}^{k} \exp\left(\mathbf{w}_j^T \mathbf{x}\right)}$$

$$= \frac{\exp\left(a_i(\mathbf{x})\right)}{\sum_{j=1}^{k} \exp\left(a_j(\mathbf{x})\right)}, \quad i = 1, \ldots, k$$

1-of-k Coding Scheme

- j-th element of $\mathbf{t}_n$:
  - 1 if $\mathbf{x}_n$ belongs to C_j
- Remaining elements are set to zero

Likelihood Function

$$p(X, T|\mathbf{w}) \quad = \quad p(X) \prod_{n=1}^{N} P(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w})$$

$$P(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}) = \prod_{j=1}^{k} P(\mathcal{C}_j|\mathbf{x}_n, \mathbf{w}_j)^{t_{nj}}$$

$$p(X, T|\mathbf{w}) = p(X) \prod_{n=1}^{N} \prod_{j=1}^{k} P(\mathcal{C}_j|\mathbf{x}_n, \mathbf{w}_j)^{t_{nj}}$$

Negative-log likelihood

$$\begin{aligned}\ell(\mathbf{w}) &= -\ln p(X) - \sum_{n=1}^{N}\sum_{j=1}^{k} t_{nj}\ln P(\mathcal{C}_j|\mathbf{x}_n, \mathbf{w}_j) \\ &= -\ln p(X) - \sum_{n=1}^{N}\sum_{j=1}^{k} t_{nj}\left[\mathbf{w}_j^T\mathbf{x}_n - \ln\left[\sum_{i=1}^{k}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)\right]\right]\end{aligned}$$

$$\nabla_{\mathbf{w}_p}\ell(W) = \sum_{n=1}^{N}\left[\frac{\exp\left(\mathbf{w}_p^T\mathbf{x}_n\right)}{\sum_{i=1}^{k}\exp\left(\mathbf{w}_i^T\mathbf{x}_n\right)} - t_{np}\right]\mathbf{x}_n,\ p = 1,\ldots,k,$$

Use gradient descent to find min of *l*, i.e. **w***