# Probabilistic Models for Sparse Regression and Classification

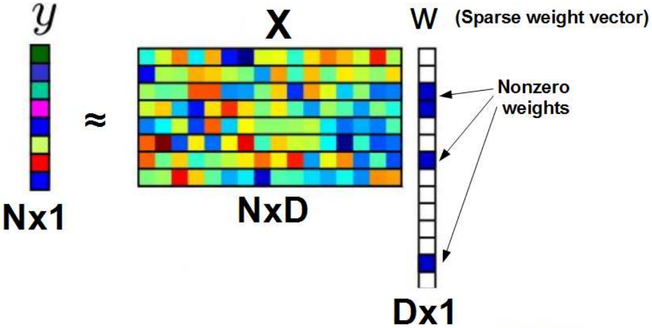
Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

March 13, 2018

# Sparse Linear Models

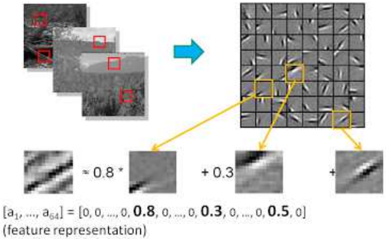
- Consider a linear (or generalized linear) model: $\boldsymbol{y} \approx \mathbf{X}\boldsymbol{w}$
- $D$ features in the data. Suppose very few are relevant for predicting $\boldsymbol{y}$
- We would excpect $\boldsymbol{w}$ to be very sparse (very few nonzeros in it)




- Very general model, appears not just in regression/classification , but also in many other problems (e.g., matrix factorization, PCA, sparse coding and dictionary learning, compressive sensing, etc.)

# Why Sparsity?

- A form of regularization. Prevents overfitting
- Fewer parameters to estimate (important when $D \gg N$)
- Can be also seen as automatically performing feature/variable selection
- Makes it easy to interpret the learned model. Also compact models (need less space to store).
- Predictions can be faster (why?)
- Many natural phenomena are inherently sparse, e.g.,
  - Natural images can be represented as sparse combination of basis images



# A Simple Bayesian Approach to Sparsity

- Recall our usual Gaussian prior on $\boldsymbol{w}$ for linear regression

$$p(\boldsymbol{w}|\alpha) = \prod_{d=1}^{D} p(w_d|\alpha) = \prod_{d=1}^{D} \left(\frac{\alpha}{2\pi}\right)^{1/2} \exp\left(-\frac{\alpha}{2} w_d^2\right)$$

- Each component of $\boldsymbol{w}$ is a zero-mean Gaussian $p(w_d|\alpha) = \mathcal{N}(w_d|0, \alpha^{-1})$
- Same inverse variance (precision) hyperparam $\alpha$ on each entry of $\boldsymbol{w}$. Can't impose sparsity on $\boldsymbol{w}$
- Let's have a separate inverse variance $\alpha_d$ for each component of $\boldsymbol{w}$

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{d=1}^{D} p(w_d|\alpha_d) = \prod_{d=1}^{D} \left(\frac{\alpha_d}{2\pi}\right)^{1/2} \exp\left(-\frac{\alpha_d}{2} w_d^2\right)$$

- We now have $D$ hyperparameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_D]$ individually controlling the variance of each component $w_d$ of $\boldsymbol{w}$. Can infer these using MLE-II, EM, etc (as we've already seen).
  - Note/Caveat: This will not usually yield an <u>exactly sparse</u> solution (i.e., won't get $w_d = 0$)

# A Hierarchical Prior to Induce Sparsity

- Our new hierarchical prior on $\boldsymbol{w}$

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{d=1}^{D} p(w_d|\alpha_d) = \prod_{d=1}^{D} \left(\frac{\alpha_d}{2\pi}\right)^{1/2} \exp\left(-\frac{\alpha_d}{2} w_d^2\right)$$

- Let's assume a gamma prior on each $\alpha_d$: $p(\alpha_d) \propto \alpha_d^{a-1} \exp^{-\alpha_d/b}$
- The marginal prior on each weight $w_d$ after integrating out $\alpha_d$

$$p(w_d) = \int p(w_d|\alpha_d)p(\alpha_d)d\alpha_d \qquad \text{(will be a Student-t distribution)}$$

- Student-t distribution is given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- Akin to penalizing $\sum_{d=1}^{D} \log |w_d|$. Leads to sparse solutions for $\boldsymbol{w}$
- This is often called "Automatic Relevance Determination" (ARD)

# Another Sparsity Promoting Prior

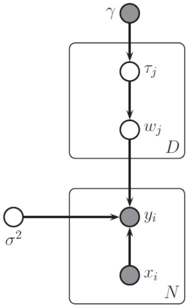
- A Laplace prior on weights can also promote sparsity (akin to putting an $\ell_1$ regularization)

$$\text{Lap}(w_j|0, 1/\gamma) = \frac{\gamma}{2} e^{-\gamma|w_j|}$$

- Such a prior (and also the previous one) is called Gaussian scale mixture prior

$$\text{Lap}(w_j|0, 1/\gamma) = \frac{\gamma}{2} e^{-\gamma|w_j|} = \int \mathcal{N}(w_j|0, \tau_j^2)\text{Ga}(\tau_j^2|1, \frac{\gamma^2}{2})d\tau_j^2$$

- Basically putting a gamma prior on the variance (a.k.a. "scale") and integrating it out



# An EM algorithm for this model..

- The joint probability distribution of data and all the unknowns

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{w}, \boldsymbol{\tau}, \sigma^2 | \mathbf{X}) &= \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N)\, \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{D}_\tau) \qquad \propto (\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2\right) |\mathbf{D}_\tau|^{-\frac{1}{2}} \\
& \quad \mathrm{IG}(\sigma^2 | a_\sigma, b_\sigma) \left[\prod_j \mathrm{Ga}(\tau_j^2 | 1, \gamma^2/2)\right] \qquad \exp\left(-\frac{1}{2}\mathbf{w}^T \mathbf{D}_\tau \mathbf{w}\right) (\sigma^2)^{-(a_\sigma + 1)} \\
& \qquad \exp(-b_\sigma/\sigma^2) \prod_j \exp\left(-\frac{\gamma^2}{2}\tau_j^2\right)
\end{aligned}
$$

- Can also optimize the above using standard optimization methods but using EM has advantages
  - Can use other types of priors on $\tau_j$ (leads to other types of sparse priors on $\boldsymbol{w}$)
  - Easily extend to finding full posterior on $\boldsymbol{w}$
  - The technique can be used in a wide variety of models that have $\ell_1$ regularization
- Treating $\tau_j$'s as latent variables and $\boldsymbol{w}$ as parameter, the M step objective is

$$\ell_c(\boldsymbol{w}) = -\frac{1}{2\sigma^2}||\boldsymbol{y} - \mathbf{X}\boldsymbol{w}||^2 - \frac{1}{2}\boldsymbol{w}^\top \mathbf{D}_\tau \boldsymbol{w} + \text{const}$$

.. easy to optimize (similar to MAP estimation for linear regression)..

# An EM algorithm for this model (contd)..

- In the E step, we infer $\tau_j$, whose posterior and the expectation is

$$p(1/\tau_j^2|\mathbf{w},\mathcal{D}) = \text{InverseGaussian}\left(\sqrt{\frac{\gamma^2}{w_j^2}}, \gamma^2\right) \qquad \mathbb{E}\left[\frac{1}{\tau_j^2}|w_j\right] = \frac{\gamma}{|w_j|}$$

$$\text{Let } \overline{\mathbf{\Lambda}} = \text{diag}(\mathbb{E}\left[1/\tau_1^2\right], \ldots, \mathbb{E}\left[1/\tau_D^2\right])$$

- Note: If we also treat $\sigma^2$ as unknown, we can infer it in the E step as

$$p(\sigma^2|\mathcal{D},\mathbf{w}) = \text{IG}(a_\sigma + (N)/2, b_\sigma + \frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})) = \text{IG}(a_N, b_N) \qquad \mathbb{E}\left[1/\sigma^2\right] = \frac{a_N}{b_N} \triangleq \overline{\omega}$$

- In the M step, we will solve the following problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmax}} -\frac{1}{2}\overline{\omega}||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 - \frac{1}{2}\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}$$

# “Hard” Sparsity

- Earlier approaches we saw had a “soft” type of sparsity ($\alpha_d/\tau_d$ tells us how important feature $d$ is)
- Consider the basic linear model $\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \epsilon$. Denote the data $(\mathbf{X}, \boldsymbol{y})$ by $\mathcal{D}$
- Total $D$ features. Suppose we have indicator variables $\gamma_d$, $d = 1, \ldots, D$
- $\gamma_d = 1$ if feature $d$ is relevant ($w_d$ is nonzero); $\gamma_d = 0$ otherwise. This is a notion of hard sparsity
- The goal is to learn $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_D\}$ (usually along with $\boldsymbol{w}$)
- If we only want $\gamma$, a natural approach would to be to do the following MAP estimation

$$\hat{\gamma} = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\mathcal{D}) = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma})p(\mathcal{D}|\boldsymbol{\gamma})$$

  where $p(\boldsymbol{\gamma})$ is the prior on $\boldsymbol{\gamma}$ and $p(\mathcal{D}|\boldsymbol{\gamma}) = \int p(\mathcal{D}|\boldsymbol{\gamma}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$ is the marginal likelihood
- A combinatorial search problem. A naïve approach to infer $\hat{\gamma}$ would require searching over a set of size $2^D$. But there are more efficient ways (e.g., greedy search)
- But first we need to write down the marginal posterior over $\boldsymbol{\gamma}$, $p(\boldsymbol{\gamma}|\mathcal{D}) \propto p(\boldsymbol{\gamma})p(\mathcal{D}|\boldsymbol{\gamma})$

# The Prior

- The marginal posterior $p(\boldsymbol{\gamma}|\mathcal{D}) \propto p(\boldsymbol{\gamma})p(\mathcal{D}|\boldsymbol{\gamma})$
- Let's consider the following prior on the binary vector $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_D\}$

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{D} \mathrm{Ber}(\gamma_j|\pi_0) = \pi_0^{||\boldsymbol{\gamma}||_0}(1-\pi_0)^{D-||\boldsymbol{\gamma}||_0}$$

- $\pi_0$ is the probability that a feature is relevant (corresponding $\gamma_d$ is nonzero)
- $||\boldsymbol{\gamma}||_0 = \sum_{d=1}^{D} \gamma_d$ is the number of nonzeros in $\boldsymbol{\gamma}$ (the $\ell_0$ norm of $\boldsymbol{\gamma}$)
- Note the form of the log-prior

$$\begin{aligned}
\log p(\boldsymbol{\gamma}|\pi_0) &= ||\boldsymbol{\gamma}||_0 \log \pi_0 + (D - ||\boldsymbol{\gamma}||_0)\log(1-\pi_0) \\
&= ||\boldsymbol{\gamma}||_0(\log \pi_0 - \log(1-\pi_0)) + \text{const} \\
&= -\lambda||\boldsymbol{\gamma}||_0 + \text{const}
\end{aligned}$$

where $\lambda \triangleq \log \frac{1-\pi_0}{\pi_0}$ controls the sparsity of the model.

# The (Marginal) Likelihood

- Let's now look at the marginal likelihood

$$p(\mathcal{D}|\gamma) = p(\mathbf{y}|\mathbf{X},\gamma) = \int\int p(\mathbf{y}|\mathbf{X},\mathbf{w},\gamma)p(\mathbf{w}|\gamma,\sigma^2)p(\sigma^2)d\mathbf{w}d\sigma^2$$

- Given $\gamma$, the prior $p(\boldsymbol{w}|\gamma,\sigma^2)$ on $\boldsymbol{w}$

$$p(w_j|\sigma^2,\gamma_j) = \begin{cases} \delta_0(w_j) & \text{if } \gamma_j = 0 \\ \mathcal{N}(w_j|0,\sigma^2\sigma_w^2) & \text{if } \gamma_j = 1 \end{cases}$$

where $w_j$ is the $j$-th element of $\boldsymbol{w}$

- This is also called a spike-and-slab prior
- Ignoring element of $\boldsymbol{w}$ that are zero, we get following marginal likelihood

$$p(\mathcal{D}|\gamma) = \int\int \mathcal{N}(\mathbf{y}|\mathbf{X}_\gamma\mathbf{w}_\gamma,\sigma^2\mathbf{I}_N)\mathcal{N}(\mathbf{w}_\gamma|\mathbf{0}_{D_\gamma},\sigma^2\sigma_w^2\mathbf{I}_{D_\gamma})p(\sigma^2)d\mathbf{w}_\gamma d\sigma^2$$

where $D_\gamma$ is the number of nonzeros in $\gamma$ and $\mathbf{X}_\gamma$ consists of the corresponding columns of $\mathbf{X}$

# The (Marginal) Likelihood

- Still have the noise variance hyperparameter $\sigma^2$
- If $\sigma^2$ is known then

$$
\begin{aligned}
p(\mathcal{D}|\gamma, \sigma^2) &= \int \mathcal{N}(\mathbf{y}|\mathbf{X}_\gamma \mathbf{w}_\gamma, \sigma^2 \mathbf{I}) \mathcal{N}(\mathbf{w}_\gamma|\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_\gamma) d\mathbf{w}_\gamma = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C}_\gamma) \\
\mathbf{C}_\gamma &\triangleq \sigma^2 \mathbf{X}_\gamma \boldsymbol{\Sigma}_\gamma \mathbf{X}_\gamma^T + \sigma^2 \mathbf{I}_N
\end{aligned}
$$

- Is $\sigma^2$ is unknown, we can put an inverse gamma proior $p(\sigma^2) = \mathrm{IG}(\sigma^2|a_\sigma, b_\sigma)$ and integrate out $\sigma^2$

$$
\begin{aligned}
p(\mathcal{D}|\gamma) &= \int\int p(\mathbf{y}|\gamma, \mathbf{w}_\gamma, \sigma^2) p(\mathbf{w}_\gamma|\gamma, \sigma^2) p(\sigma^2) d\mathbf{w}_\gamma d\sigma^2 \\
&\propto |\mathbf{X}_\gamma^T \mathbf{X}_\gamma + \boldsymbol{\Sigma}_\gamma^{-1}|^{-\frac{1}{2}} |\boldsymbol{\Sigma}_\gamma|^{-\frac{1}{2}} (2b_\sigma + S(\gamma))^{-(2a_\sigma + N - 1)/2}
\end{aligned}
$$

$S(\gamma)$ denotes the residual sum-of-squares $\boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbf{X}_\gamma (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma + \boldsymbol{\Sigma}_\gamma^{-1})^{-1} \mathbf{X}_\gamma^\top \boldsymbol{y}$

# Solving for $\gamma$

- Now we have an expression (exact/approximate) for (log)$p(\mathcal{D}|\gamma)$
- This combined with $\log p(\gamma) \propto -\lambda||\gamma||_0$ gives the posterior over $\gamma$
- Now the goal is to find $\gamma$
- One way is to do it by greedily choosing the next feature to add
- Some heuristics
  - Single best replacement (start with $\gamma = 0$ and consider all one-bit flips)
  - Matching pursuit (for ordinary least squares problems)

$$j^* = \arg\min_{j \notin \gamma_t} \min_{\beta} ||\mathbf{y} - \mathbf{X}\mathbf{w}_t - \beta \mathbf{x}_{:,j}||^2$$

$$\beta = \mathbf{x}_{:,j}^T \mathbf{r}_t / ||\mathbf{x}_{:,j}||^2$$

$$j^* = \arg\max \mathbf{x}_{:,j}^T \mathbf{r}_t$$

  where $\boldsymbol{r}_t = ||\boldsymbol{y} - \mathbf{X}\boldsymbol{w}_t||$ is the current residual. Thus we choose the feature (a column in the $N \times D$ feature matrix $\mathbf{X}$) that has the maximum correlation with current residual

- Do also do fully Bayesian inference using MCMC or VB