Single-Parameter Models (Contd.)

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 11, 2018

Recap

- A probabilistic model consists of the observations $\boldsymbol{y}$ and one or more unknowns θ
- Assuming the unknowns as random variables, can define a joint distribution $p(\boldsymbol{y}, \theta)$ over $\boldsymbol{y}$ and θ
- Apply the conditioning formula (Bayes rule) to infer the posterior distribution over the unknowns

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \theta)}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}} \propto \text{Likelihood} \times \text{Prior}$$

where marginal likelihood $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\theta)p(\theta)d\theta = \mathbb{E}_{p(\theta)}[p(\boldsymbol{y}|\theta)]$ (also called "model evidence")

- The posterior predictive distribution for new data $\boldsymbol{y}_*$ conditioned on the past data $\boldsymbol{y}$

$$\underbrace{p(\boldsymbol{y}_*|\boldsymbol{y})}_{\text{posterior predictive}} = \int \underbrace{p(\boldsymbol{y}_*|\theta)}_{\text{point predictive}} p(\theta|\boldsymbol{y})d\theta = \mathbb{E}_{p(\theta|\boldsymbol{y})}[p(\boldsymbol{y}_*|\theta)]$$

.. i.e., posterior-weighted average of point predictive distributions $p(\boldsymbol{y}_*|\theta)$ over all values of θ

- Note: Even if it is not explicit in the notation, the terms such as $p(\boldsymbol{y}|\theta)$, $p(\theta)$, $p(\boldsymbol{y})$ are usually conditioned on other things too, e.g, the model index and their respective fixed hyperparameters

Recap

- Instead of inferring the full posterior, another option is to do point estimation of parameters
  - Cheaper alternative of full posterior inference, but doesn't capture the uncertainty in θ
- Examples: Maximum likelihood estimation (MLE) and Maximum-a-Posteriori (MAP) estimation
- Given N i.i.d. observations $\boldsymbol{y} = \{y_1, \ldots, y_N\}$ from a model $p(\boldsymbol{x}|\theta)$
  - MLE: Maximizes the (log of) likelihood $p(\boldsymbol{y}|\theta)$ w.r.t. θ

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log p(\boldsymbol{y}|\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(y_n|\theta)$$

  - MAP: Maximizes the (log of) posterior $p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta)p(\theta)$ w.r.t. θ

$$\hat{\theta}_{MAP} = \arg\max_{\theta} [\log p(\boldsymbol{y}|\theta) + \log p(\theta)] = \arg\max_{\theta} \sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta)$$

- Loss function view: Negative log-lik. (NLL) = loss function. Negative log-prior = regularizer
- With point estimation, we cannot get posterior predictive $p(\boldsymbol{y}_*|\boldsymbol{y})$ but only point predictive $p(\boldsymbol{y}_*|\hat{\theta})$

Recap

- Coin toss example. Each likelihood term is Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Assume Beta prior on θ: $p(\theta) = \text{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, α,β are fixed hyperparams
- The posterior distribution will be proportional to the product of likelihood and prior

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- The above posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n, \beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n)$
- Conjugate prior: If posterior has the same form as the prior (which is the case above - both Beta)
- The probability of next toss being head

$$
p(\boldsymbol{x}_{N+1}=1|\mathbf{X}) = \int_0^1 P(\boldsymbol{x}_{N+1}=1|\theta)p(\theta|\mathbf{X})d\theta = \int_0^1 \theta\times\text{Beta}(\theta|\alpha+N_1,\beta+N_0)d\theta = \mathbb{E}_{p(\theta|\mathbf{X})}[\theta] = \frac{\alpha+N_1}{\alpha+\beta+N}
$$

- Thus the posterior predictive distribution $p(\boldsymbol{x}_{N+1}|\mathbf{X}) = \text{Bernoulli}(\mathbb{E}[\theta|\mathbf{X}])$

Single-Parameter Model: Gaussian with Known Variance

- Consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance σ^2 to be known/fixed
- Let's estimate μ given the data $\mathbf{X}$ using fully Bayesian inference (not MLE/MAP)
- We first need a prior distribution for the unknown param. μ
- Let's choose a Gaussian prior on μ, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with μ_0, σ_0^2 as fixed
- Therefore this is also a single-parameter model (only μ is the unknown)

Single-Parameter Models: Gaussian with Known Variance

- The posterior distribution for the unknown mean parameter μ

$$p(\mu|\mathbf{X}) = \frac{p(\mathbf{X}|\mu)p(\mu)}{p(\mathbf{X})} \quad \propto \quad \prod_{n=1}^{N} \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

- Simplifying the above (using completing the squares trick) gives $p(\mu|\mathbf{X}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$ with

$$\begin{aligned}\frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\\ \mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x} \qquad \left(\text{where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}\right)\end{aligned}$$

- Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)
- Consider what happens as N (number of observations) grows very large?
  - The posterior's variance σ_N^2 approaches σ^2/N (and goes to 0 as $N \to \infty$)
  - The posterior's mean μ_N approaches $\bar{x}$ (which is also the MLE solution)

Single-Parameter Models: Gaussian with Known Variance

- What is the posterior predictive distribution $p(x_*|\mathbf{X})$ of a new observation x_*?
- Using the inferred posterior $p(\mu|\mathbf{X})$, we can find the posterior predictive distribution

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu,\sigma^2)p(\mu|\mathbf{X})d\mu = \int \mathcal{N}(x_*|\mu,\sigma^2)\mathcal{N}(\mu|\mu_N,\sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N,\sigma^2+\sigma_N^2)$$

- Note that, as per the above, the uncertainty in distribution of x_* now has two components
  - σ^2: Due to the noisy observation model
  - σ_N^2: Due to the uncertainty in the estimate of the mean μ
- In contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

$$p(x_*|\mathbf{X}) = \int p(x_*|\mu,\sigma^2)p(\mu|\mathbf{X})d\mu \approx p(x_*|\hat{\mu},\sigma^2) = \mathcal{N}(x_*|\hat{\mu},\sigma^2)$$

.. which doesn't incorporate the uncertainty in our estimate of μ (since we used a point estimate)

- Note that as $N \to \infty$, both approaches would give the same $p(x_*|\mathbf{X})$ since $\sigma_N^2 \to 0$

Single-Parameter Models: Gaussian with Known Mean

- Again consider N i.i.d. observations $\mathbf{X} = \{x_1, \dots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean μ to be known/fixed
- Let's estimate σ^2 given the data $\mathbf{X}$ using fully Bayesian inference (not MLE/MAP)
- We first need a prior distribution for σ^2. What prior $p(\sigma^2)$ to choose in this case?
- If we want a conjugate prior, it should have the form form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

- An inverse-gamma prior $IG(\alpha, \beta)$ has this form (α, β are shape and scale hyperparams, resp)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right] \qquad \text{(note: mean of } IG(\alpha, \beta) = \frac{\beta}{\alpha - 1}\text{)}$$

- (Verify) The posterior $p(\sigma^2|\mathbf{X}) = IG(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^{N}(x_n - \mu)^2}{2})$. Again IG due to conjugacy.

Working with Gaussians: Variance vs Precision

- Often, it is easier to work with the precision (=1/variance) rather than variance

$$p(x_n|\mu,\tau) = \mathcal{N}(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(x_n-\mu)^2\right]$$

- If mean is known, for precision Gamma(α,β) is a conjugate prior to Gaussian likelihood

$$p(\tau) \propto (\tau)^{(\alpha-1)} \exp\left[-\beta\tau\right] \qquad \text{(note: mean of Gamma}(\alpha,\beta) = \frac{\alpha}{\beta})$$

.. where α and β are the shape and rate hyperparamers, respectively, for the Gamma

- (Verify) The posterior $p(\tau|\mathbf{X})$ will also be Gamma$(\alpha + \frac{N}{2}, \beta + \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2})1$
- Note: Gamma distribution can be defined in terms of shape and scale or shape and rate parametrization (scale = 1/rate). Likewise, inverse Gamma can also be defined both shape and scale (which we saw) as well as shape and rate parametrizations.

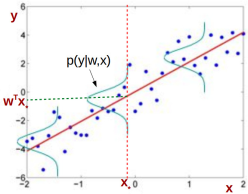
Another Single-Parameter Model

Probabilistic Linear Regression

(with fixed hyperparamers)

Linear Regression

- Given: Data $\mathcal{D}$ with N training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, features: $\boldsymbol{x}_n \in \mathbb{R}^D$, response $y_n \in \mathbb{R}$
- $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]^\top$: $N \times D$ feature matrix, $\boldsymbol{y} = [y_1 \dots y_N]^\top$: $N \times 1$ resp. vector
- Assume a "noisy" linear model with regression weight vector $\boldsymbol{w} \in \mathbb{R}^D$: $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$
- Gaussian noise: $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$, β: precision (inverse variance) of Gaussian
- Thus each response y_n also has a Gaussian distribution: $P(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$

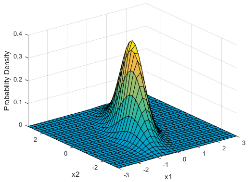


- Goal: Learn regression weight vector $\boldsymbol{w}$ (this is our unknown θ) to predict y_* for a new $\boldsymbol{x}_*$

Multivariate Gaussian Distribution

- Distribution over a multivariate r.v. vector $\mathbf{x} \in \mathbb{R}^D$ of real numbers
- Defined by a **mean vector** $\boldsymbol{\mu} \in \mathbb{R}^D$ and a $D \times D$ **cov. matrix** $\boldsymbol{\Sigma}$ (its inverse is precision matrix)

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



- The covariance matrix $\boldsymbol{\Sigma}$ must be symmetric and positive definite
  - All eigenvalues are positive
  - $\boldsymbol{z}^\top \boldsymbol{\Sigma} \boldsymbol{z} > 0$ for any real vector $\boldsymbol{z}$

Likelihood and Prior

- Assuming i.i.d. observations, the likelihood model $p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{w}, \boldsymbol{x}_n)$
- Note: Here data $\mathcal{D}$ refers to $(\mathbf{X}, \boldsymbol{y})$ but input $\mathbf{X}$ is not being modeled; only $\boldsymbol{y}$ is being modeled
- As we saw, each likelihood term is a univariate Gaussian: $p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$
- Collectively, the overall likelihood $p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X})$ will be an N dimensional multivariate Gaussian

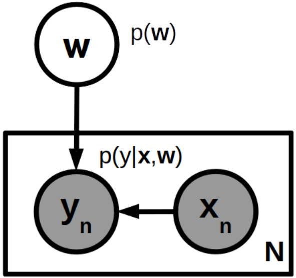
$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1}) = \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N)$$

- The unknown parameter $\boldsymbol{w} \in \mathbb{R}^D$. Can assume a zero mean D-dim multivar. Gaussian prior

$$p(\boldsymbol{w}|\lambda) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)$$

  where λ denotes the precision of the Gaussian (how shrunk it is towards the mean)
- The Gaussian prior $p(\boldsymbol{w}|\lambda) \propto \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}\right]$ is akin to imposing an ℓ_2 regularizer (and λ is like the regularization constant)

The Graphical Model Notation



(Hyperparameters not shown as they are fixed/known)

The Posterior

- The posterior over the weight vector $\boldsymbol{w}$ (for now, we will assume hyperparams β and λ to be fixed)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \frac{P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta)}{P(\boldsymbol{y}|\mathbf{X}, \beta, \lambda)}$$

- Computing $P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ (like Bernoulli-Beta case, doing it only upto proportionality constant)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta)$$

- After some algebra, this gets simplified into the following (proof on the next two slides)

$$\begin{aligned} P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y} \end{aligned}$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$, we will get

$$
\begin{aligned}
p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) \quad &= \quad \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\
&\propto \quad \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \\
&= \quad \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{\beta}{2}(\boldsymbol{y}^\top\boldsymbol{y} + \boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&\propto \quad \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{\beta}{2}(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&= \quad \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D + \beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w} - 2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y}\right)\right]
\end{aligned}
$$

- We will now try to bring the exponent into a quadratic form to see if it corresponds to some Gaussian. So basically, we will use the "complete the square" trick

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of a Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]$$

- Let's multiply and divide $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$ by $\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$
- This gives the following up to a prop. constant (remember $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is constant w.r.t. $\boldsymbol{w}$):

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]$$

- Finally comparing with the expression of $\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can see that

$$\begin{aligned} \boldsymbol{\Sigma} &= (\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \beta \mathbf{X}^\top \boldsymbol{y} \quad \Rightarrow \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) \end{aligned}$$

- Note: The above expression for the posterior can also be directly obtained using properties of Gaussian distributions (we will see those in the coming lectures)

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$
\begin{aligned}
P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

- Note the form of the mean $\boldsymbol{\mu}$ of the posterior. Equivalent to the solution of ridge regression.
  - Also, if we do a direct MAP estimation for $\boldsymbol{w}$, we will have $\boldsymbol{w}_{MAP} = \boldsymbol{\mu}$
- Setting λ to 0 makes μ equivalent to the MLE solution
- However, MLE and MAP are only point estimates whereas we now have the full posterior over $\boldsymbol{w}$
- Therefore, the solution given by the full posterior in a way subsumes MAP/MLE solutions

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)
- The posterior predictive distribution in this case will be

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) d\boldsymbol{w}$$

- The result and therefore the posterior predictive will be another Gaussian (we will see proof later)

$$\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma} \boldsymbol{x}_*)}$$

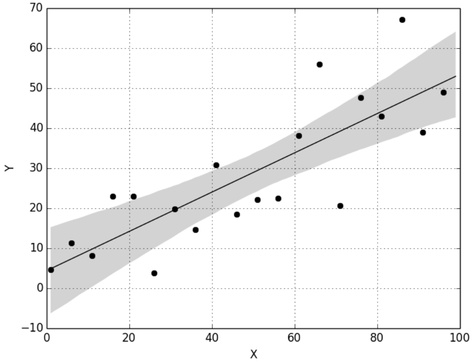
- In contrast, MLE and MAP make "plug-in" predictions (using the point estimate of $\boldsymbol{w}$)

$$\begin{aligned} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MLE}) &= \mathcal{N}(\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MLE prediction} \\ p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MAP}) &= \mathcal{N}(\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MAP prediction} \end{aligned}$$

- Important: Unlike MLE/MAP, the variance of y_* also depends on the input $\boldsymbol{x}_*$ (this, as we will see later, will be very useful in sequential decision-making problems such as active learning)

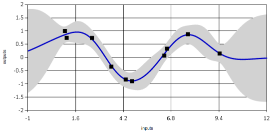
Posterior Predictive Distribution: An Illustration

Black dots are training examples



Regions with more training examples have smaller predictive variance

Nonlinear Regression?



- Can extend the linear regression model to handle nonlinear regression problems
- One way is to replace the feature vectors $\boldsymbol{x}$ by a nonlinear mapping $\phi(\boldsymbol{x})$

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \phi(\boldsymbol{x}), \beta^{-1})$$

- The nonlinear mapping can be defined directly, e.g., for a one-dimensional feature x

$$\phi(x) = [1, x, x^2]$$

- Alternatively, a kernel function can be used to implicitly define the nonlinear mapping
- More on nonlinear regression when we discuss Gaussian Processes

What about the hyperparameters of the regression model?

- If hyperparameters are to be estimated, we will have a hierarchical/multiparameter model
- Posterior inference in slighltly more involved in this case
- Iterative methods required to learn the weight vector and the hyperparameters, e.g.,
  - Marginal likelihood maximization for hyperparameter estimation
  - Expectation maximization (EM)
  - MCMC or variational inference
- More on this later..